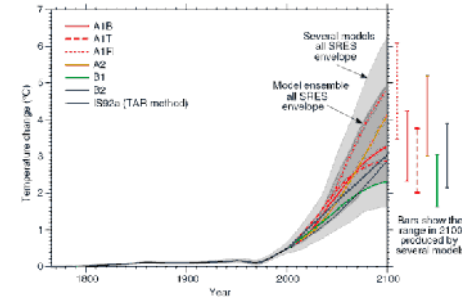
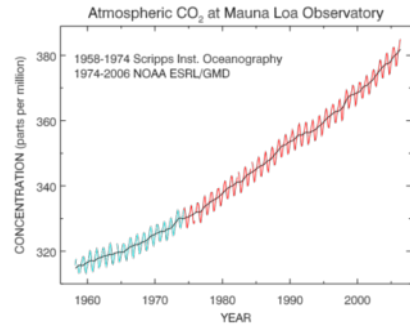


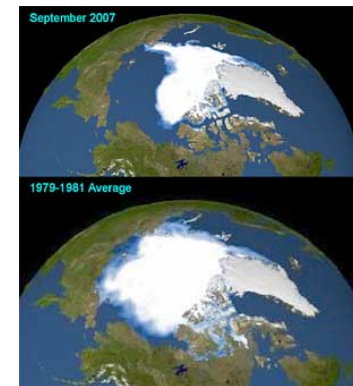
Climate Change: The defining issue of our era

- **The planet is warming**
 - Multiple lines of evidence
 - Credible link to human GHG (green house gas) emissions
- **Consequences can be dire**
 - Extreme weather events, regional climate and ecosystem shifts, abrupt climate change, stress on key resources and critical infrastructures
- **There is an urgency to act**
 - Adaptation: "Manage the unavoidable"
 - Mitigation: "Avoid the unmanageable"
- **The societal cost of both action and inaction is large**



Russia Burns, Moscow Chokes

NATIONAL GEOGRAPHIC, 2010



The Vanishing of the Arctic Ice cap

ecology.com, 2008

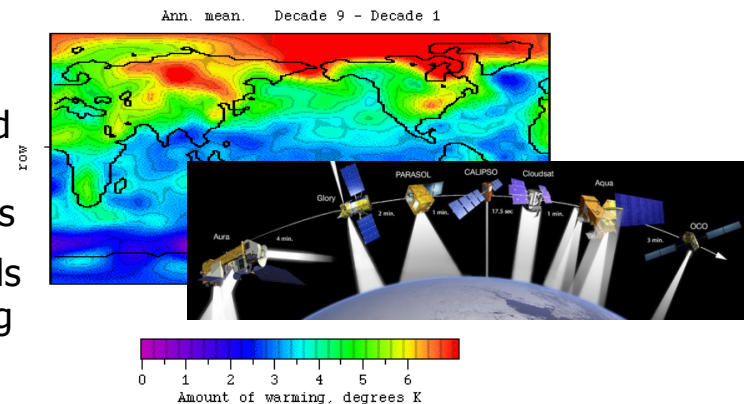
Key outstanding science challenge:

Actionable predictive insights to credibly inform policy

Data-Driven Knowledge Discovery in Climate Science

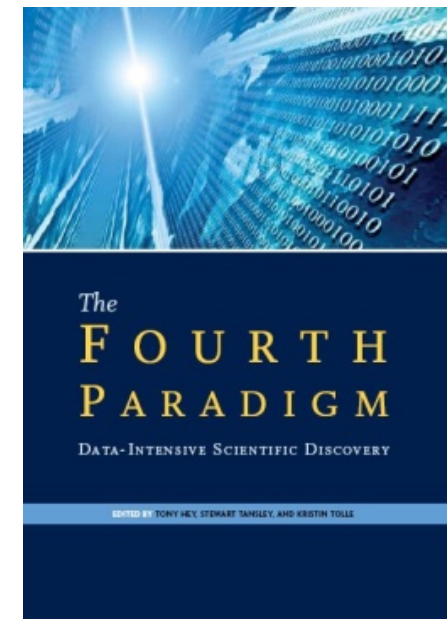
- **From data-poor to data-rich transformation**

- **Sensor Observations:** Remote sensors like satellites and weather radars as well as in-situ sensors and sensor networks like weather station and radiation measurements
- **Model Simulations:** IPCC climate or earth system models as well as regional models of climate and hydrology, along with observed data based model reconstructions



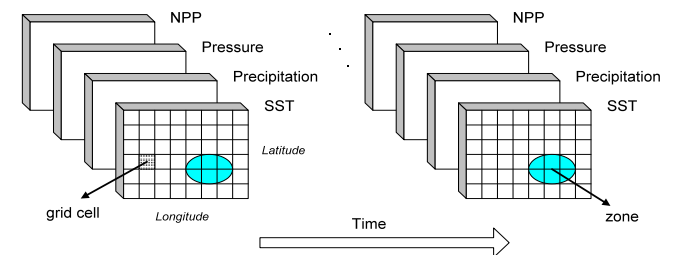
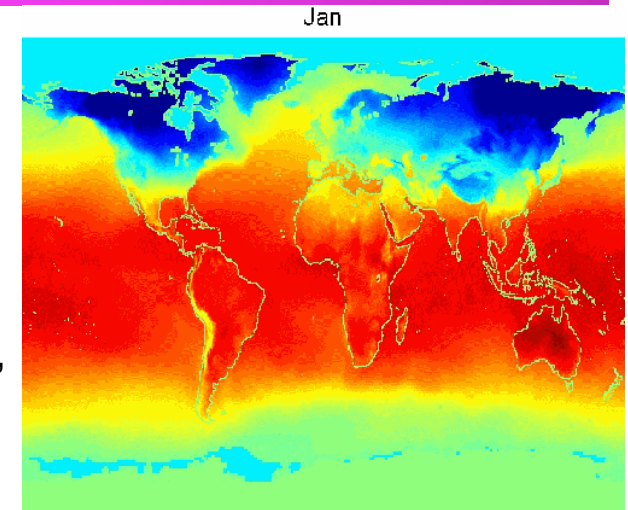
- Data guided processes can complement hypothesis guided data analysis to develop predictive insights for use by climate scientists, policy makers and community at large.

"The world of science has changed ... data-intensive science [is] so different that it is worth distinguishing [it] ... as a new, fourth paradigm for scientific exploration." - Jim Gray



Data Mining Challenges

- **Spatio-temporal nature of data**
 - spatial and temporal autocorrelation.
 - Multi-scale/Multi-resolution nature
- **Scalability**
 - Size of Earth Science data sets can be very large, For example, for each time instance,
 - ◆ $2.5^\circ \times 2.5^\circ$: 10K locations for the globe
 - ◆ $250\text{m} \times 250\text{m}$: ~10 billion
 - ◆ $50\text{m} \times 50\text{m}$: ~250 billion
- **High-dimensionality**
- **Noise and missing values**
- **Long-range spatial dependence**
- **Long memory temporal processes**
- **Nonlinear processes, Non-Stationarity**
- **Fusing multiple sources of data**



Illustrative Applications of Data Mining

- Monitoring of global forest cover
- Understanding the impact of climate change using data driven analysis.

Monitoring Forest Cover Change: Motivation

- Forests are a critical component of planet.
 - Act as sink of carbon from the atmosphere.
 - Provide ecological diversity and protect soil.
 - Livelihood for millions of people.
- Massive degradation in forest cover due to logging, conversions to cropland or plantations and natural disasters like fires.
- Quantifiable knowledge about changes in forest cover is critical for effective management of forest resources
 - Carbon Trading
 - UN REDD: monetary payments for preservation of forests.



Purveyors of water, consumers of carbon, treasure-houses of species, the world's forests are ecological miracles. They must not be allowed to vanish. The Economist, Sep 23rd 2010



Illegal deforestation in Para, Brazil. Source: Greenpeace

State of the Art in Land cover change detection

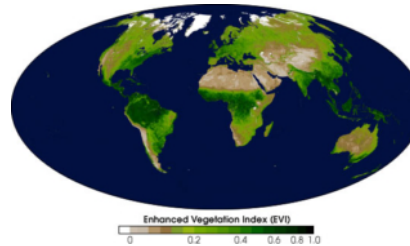
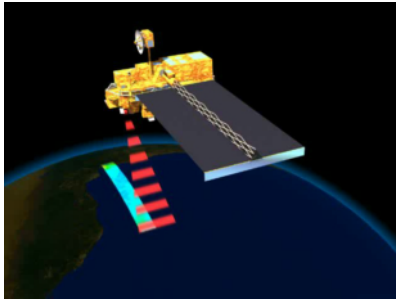
- Primarily based on examining differences between two or more high quality satellite images acquired on different dates.

Limitations:

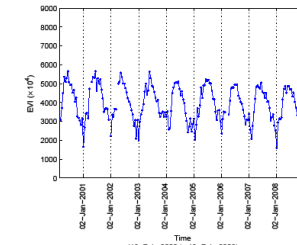
- High quality observations are infrequent in many parts of the world such as the tropics.
- Unable to detect changes outside the image acquisition window.
- Difficult to identify when the change has occurred.
- Parameters such as rate of change, extent, speed, and pattern of growth cannot be derived.
- Requires training data for each specific change of interest making it inherently unsuitable for global analysis.



Alternate approach: Analyzing Vegetation Time Series



EVI shows density of plant growth on the globe.



EVI time series for a location

- Daily Remote Sensing observations are available from MODIS aboard AQUA and TERRA satellites.

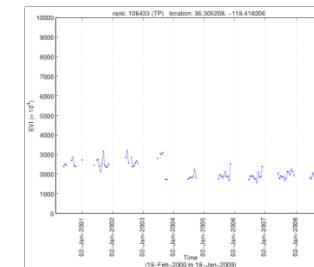
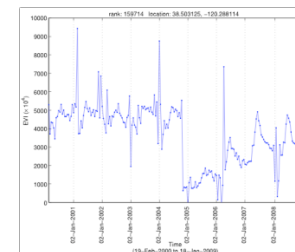
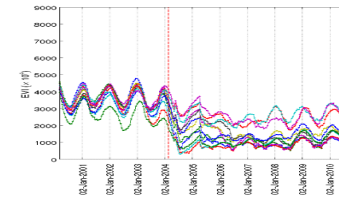
- High temporal frequency (daily for multi-spectral data and bi-weekly for the Vegetation index products like EVI, FPAR)

- Time series based approaches can be used for

- Detection of a greater variety of changes.
 - Identifying when the change occurred
 - Characterization of the type of change eg. abrupt vs gradual
 - Near-real time change identification

- Challenges

- Poor data quality and high variability
 - Coarse spatial resolution of observations (250 m)
 - massive data sets: 10 billion locations for the globe



Previous work: Time Series Change Detection

Time series change detection problem has been addressed in a variety of fields under different names:

E.g. statistical process control, curve segmentation (computer graphics & vision), segmented regression

- Statistics
- Signal processing
- Control theory
- Industrial process control
- Computer graphics & vision (curve segmentation)
- Network Intrusion Detection
- Fraud Detection (telecommunications, etc.)
- Health Care (Statistical Surveillance)
- Industrial Processes (process control and quality control)
- **Land Cover Change**

- Parameter Change
CUSUM-type approaches, Page [1957], Chernoff and Zacks [1964], Picard [1985]
- Segmentation
Linear Model: Himberg et al. [2001], Keogh et al. [2001], Hawkins and Merriam [1973]
Polynomial Model: Guralnik and Srivastava [1999]
Wavelet Model: Sharifzadeh et al. [2005]
- Predictive
Ge and Smyth [2000], Roy, Jin, Lewis and Justice [2005]
- Subspace Approach
Moskvina and Zhigljavsky [2003]
- Anomaly Detection
Chan and Mahoney [2005], Yamanishi and Takeuchi [2002], Ide and Kashima [2004], Chandola, Banerjee and Kumar [2008]

Novel Time Series Change Detection Techniques

Existing Time series change detection algorithms do not address unique characteristics of eco-system data like noise, missing values, outliers, high degree of variability (across regions, vegetation types, and time).

Segmentation based approaches

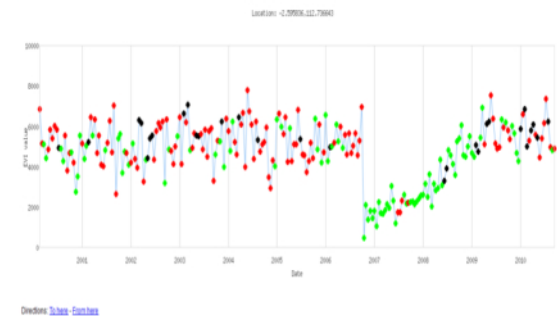
- Divide time series into homogenous segments.
- Boundary of segments become the change points.
- Useful for detection land cover conversions like forest to cropland, etc.

Prediction based approaches

- Build a prediction model for the location using previous observations.
- Use the deviation of subsequent observations from the predicted value by the model to identify changes/disturbances.
- Useful for detecting deviations from the normal vegetation model.



EVI time series for a 250 m by 250 m of land in Iowa, USA that changed from fallow land to agriculture land.



FPAR time series for a forest fire location in California, USA.

- S. Boriah, V. Kumar, M. Steinbach, et al., *Land cover change detection: a case study*, KDD 2008.
- V. Mithal, S. Boriah, A. Garg, M. Steinbach, V. Kumar et al., *Monitoring global forest cover using data mining*. ACM Transactions on Intelligent Systems and Technology, 2011 (In Press)

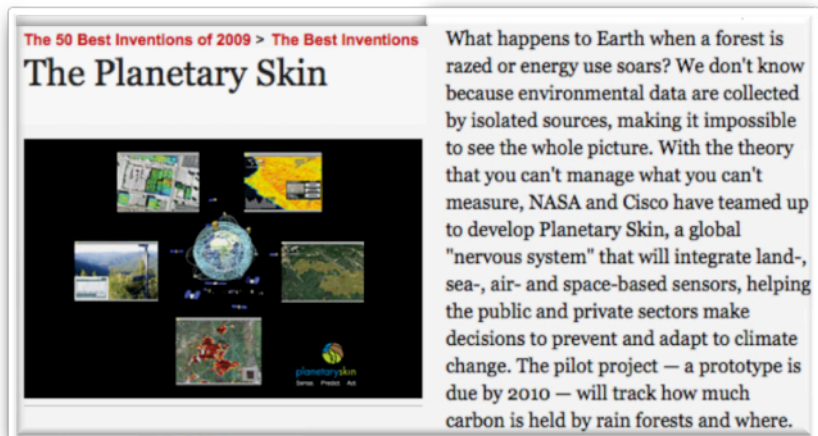
Monitoring of Global Forest Cover

- Automated Land change Evaluation, Reporting and Tracking System (ALERT)
 - Planetary Information System for assessment of ecosystem disturbances:
 - Forest fires, droughts, floods, logging/deforestation, conversion to agriculture
- This system will help
 - quantify the carbon impact of these changes
 - Understand the relationship to global climate variability and human activity
- Provide **ubiquitous web-based access** to changes occurring across the globe, creating public awareness



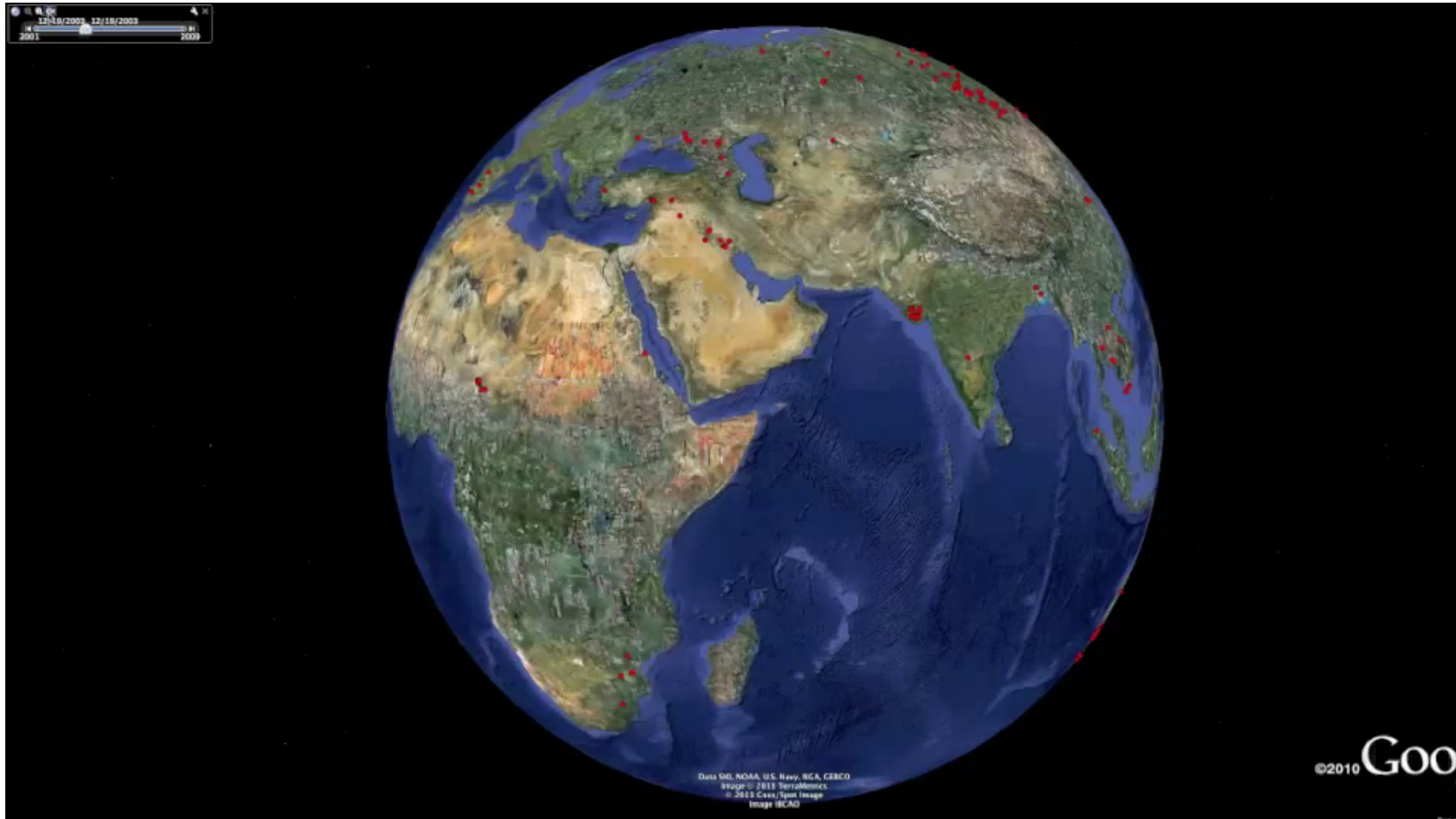
TIME

The 50 Best Inventions of 2009



Case Study 1:

Monitoring Global Forest Cover

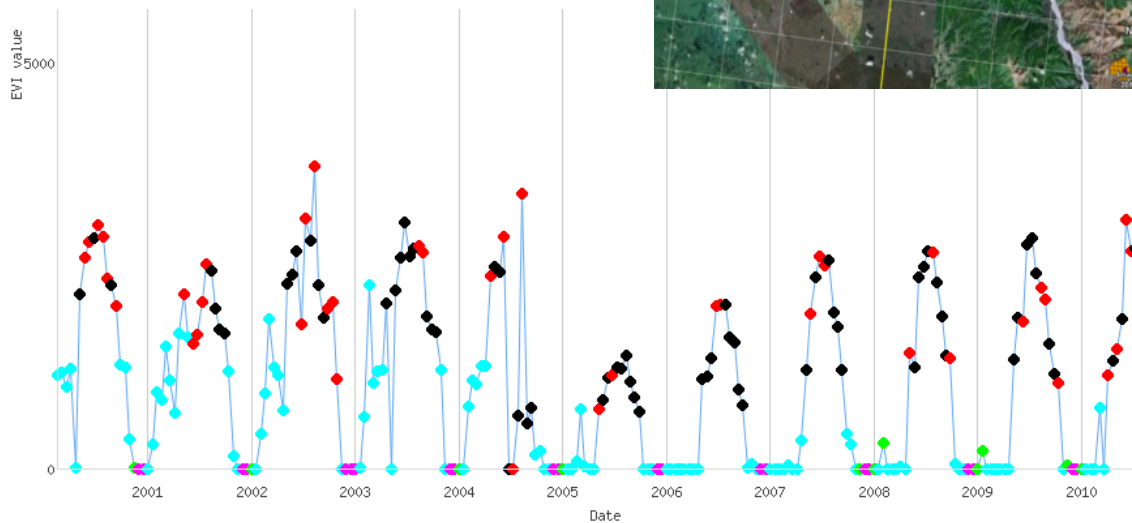
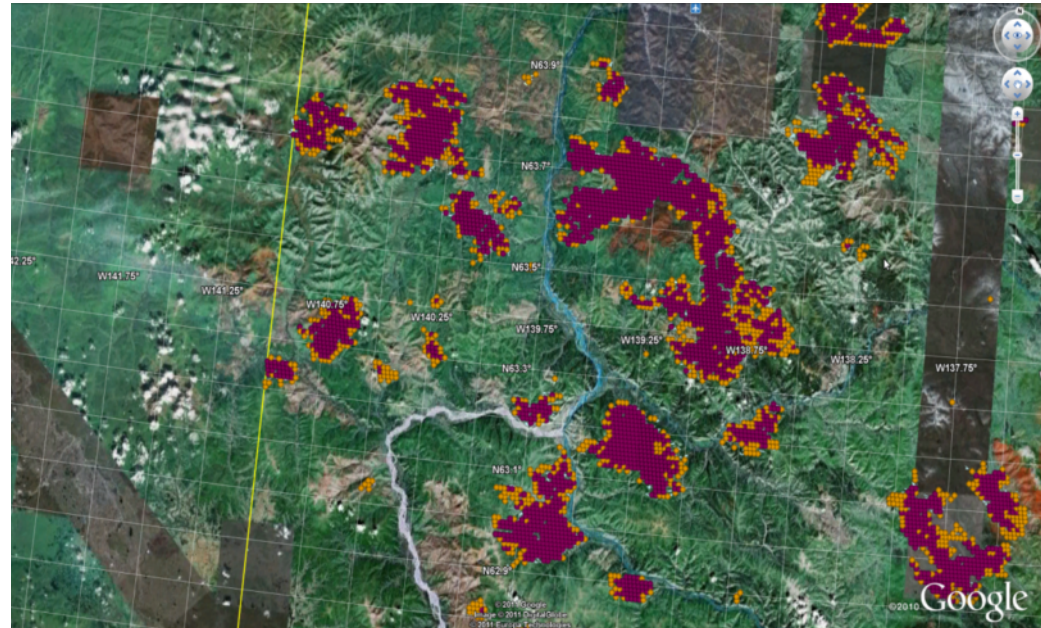


Fires in Northern Latitude (Canada/Russia) 2001-2009



Forest Fires in Canada

Massive Fires in Canada have converted the forests into source of carbon in the atmosphere.



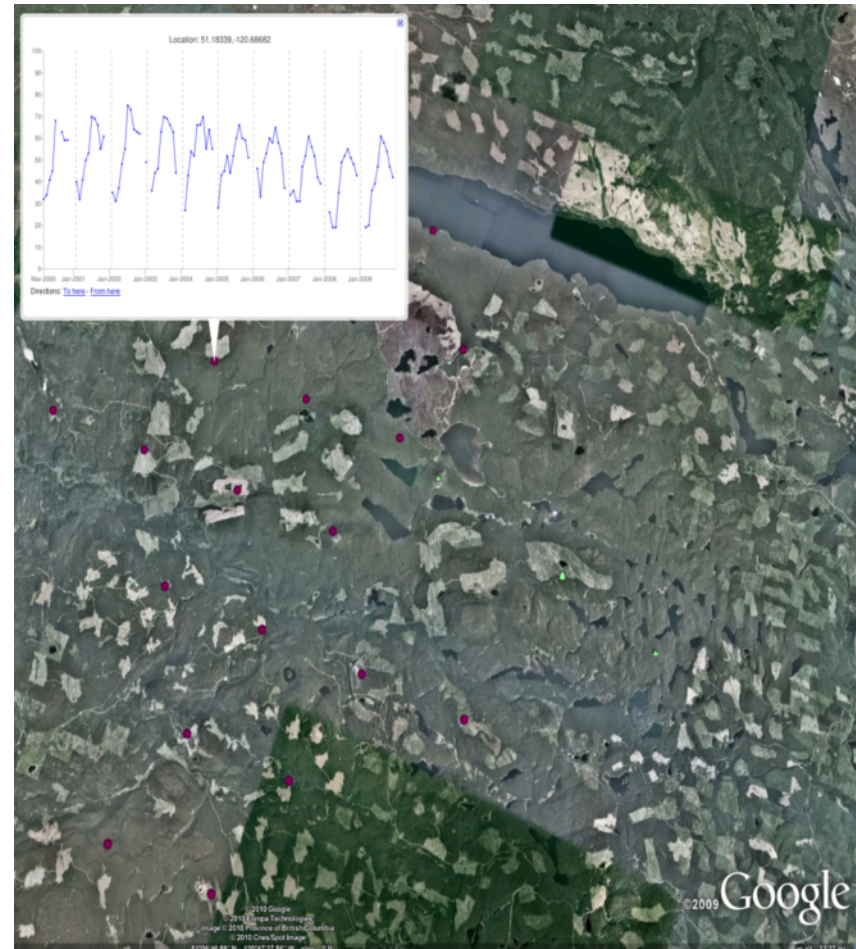
Logging in Canada



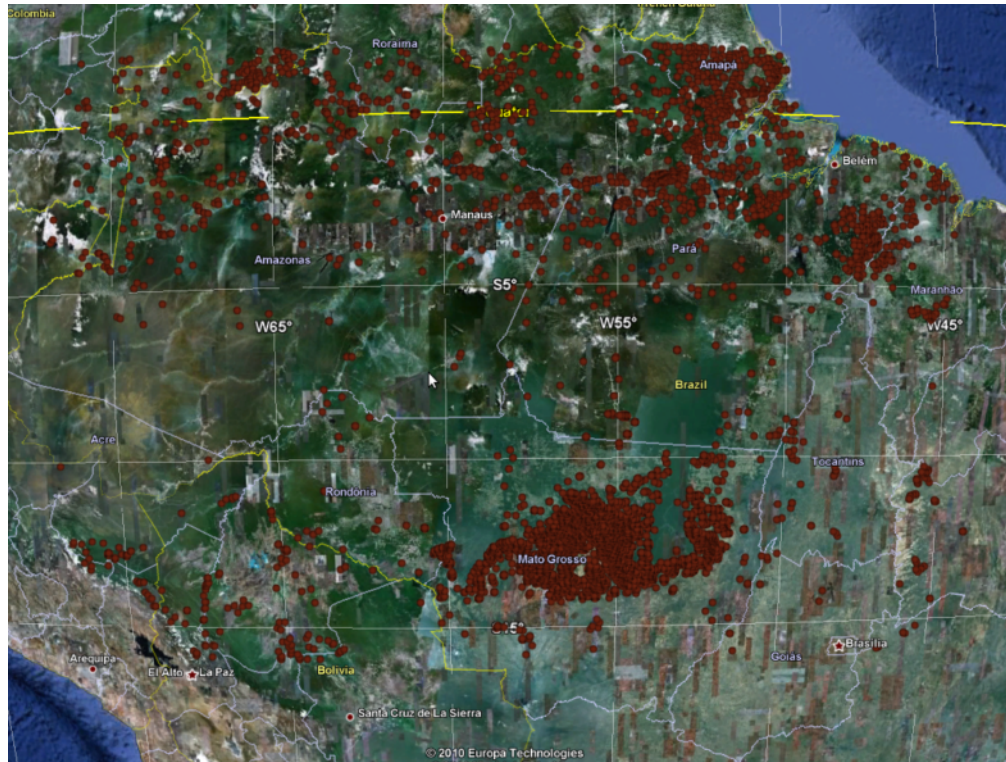
- Logging has produced clear cut areas in British Columbia, which can be identified as regular, generally rectangular shapes.

- The highly reflective clear cut areas stand out in marked contrast to the dark green forested areas.

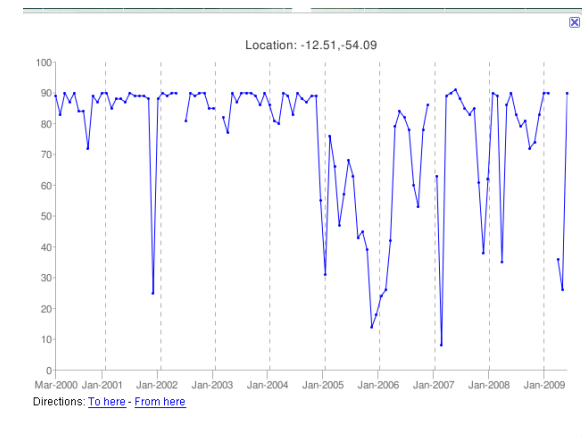
(Source: NASA)



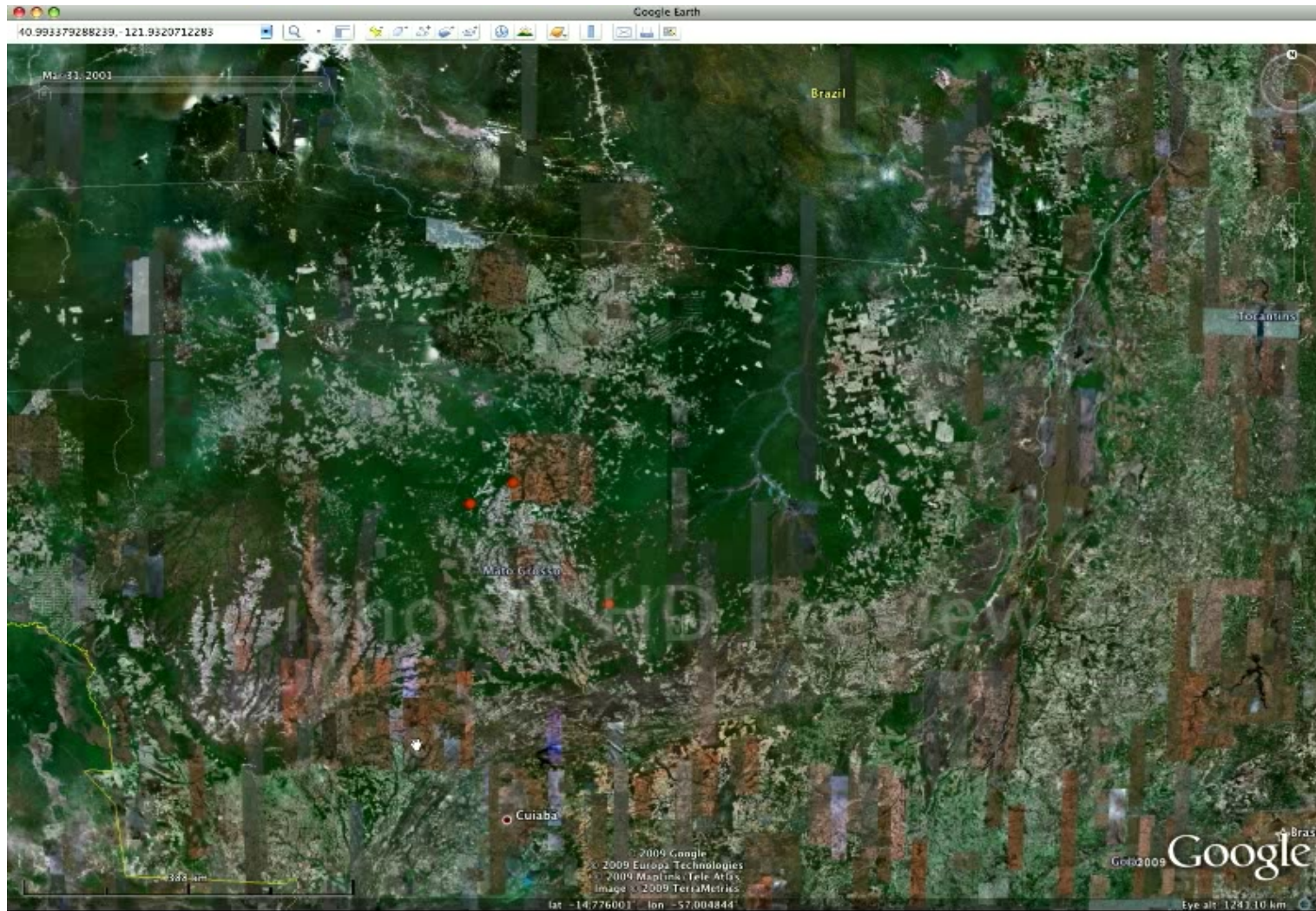
Deforestation in the Amazon Rainforest



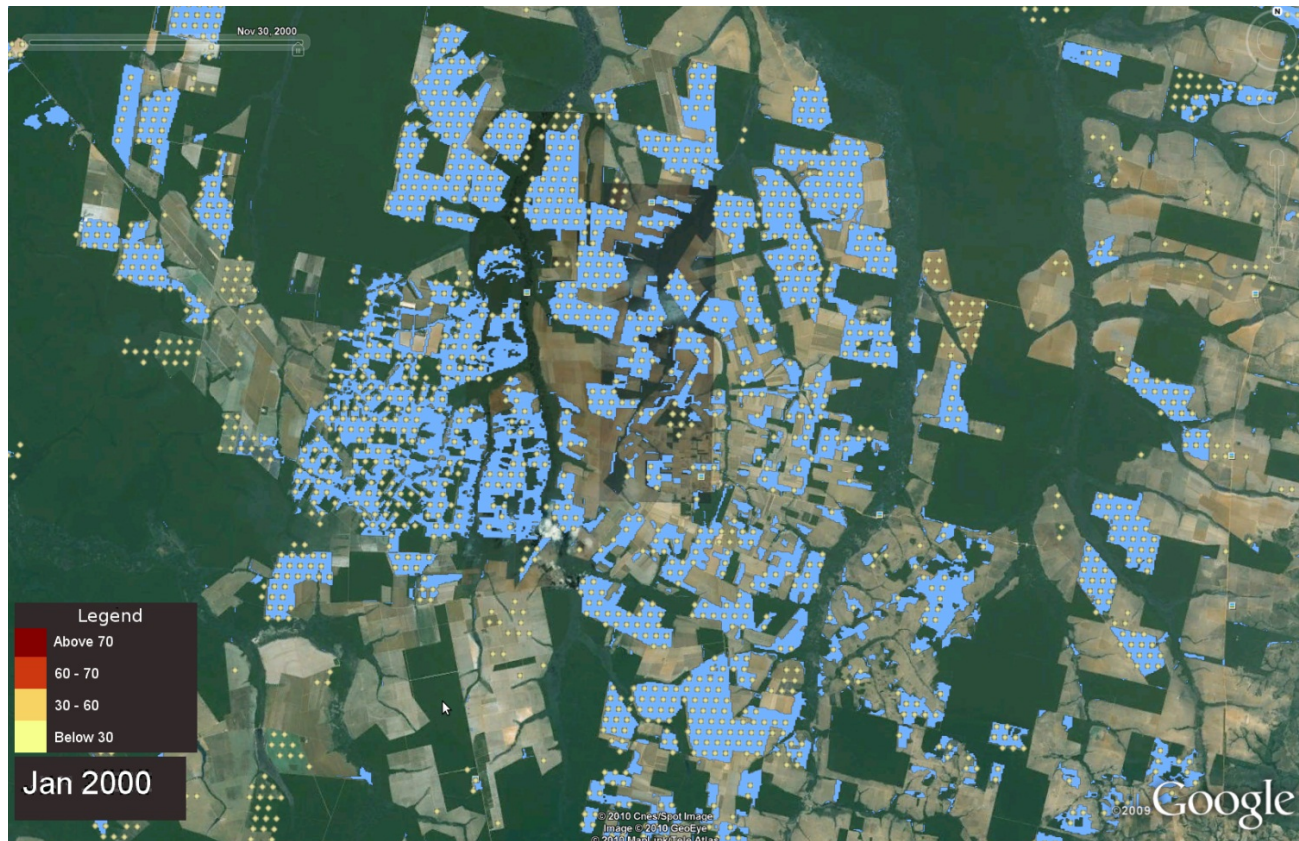
Brazil Accounts for almost 50% of all humid tropical forest clearing, nearly 4 times that of the next highest country, which accounts for 12.8% of the total.



Amazon Deforestation Animation 2001-2009



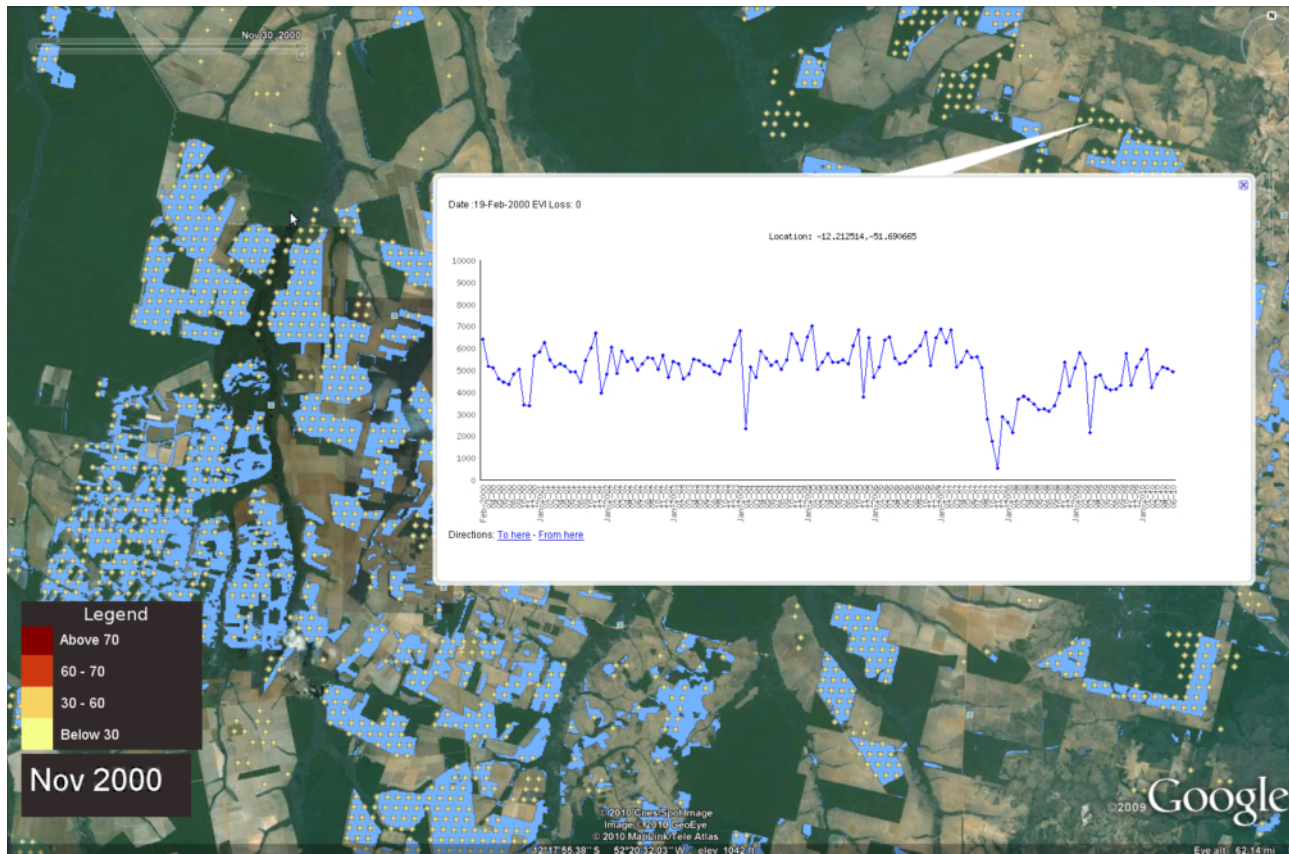
Deforestation in the Amazon Rainforest: Comparison with PRODES



PRODES is a system for monitoring deforestation in Brazilian Amazon.

The blue polygons are deforestation changes marked by PRODES.
Yellow dots are events detected by our algorithm.

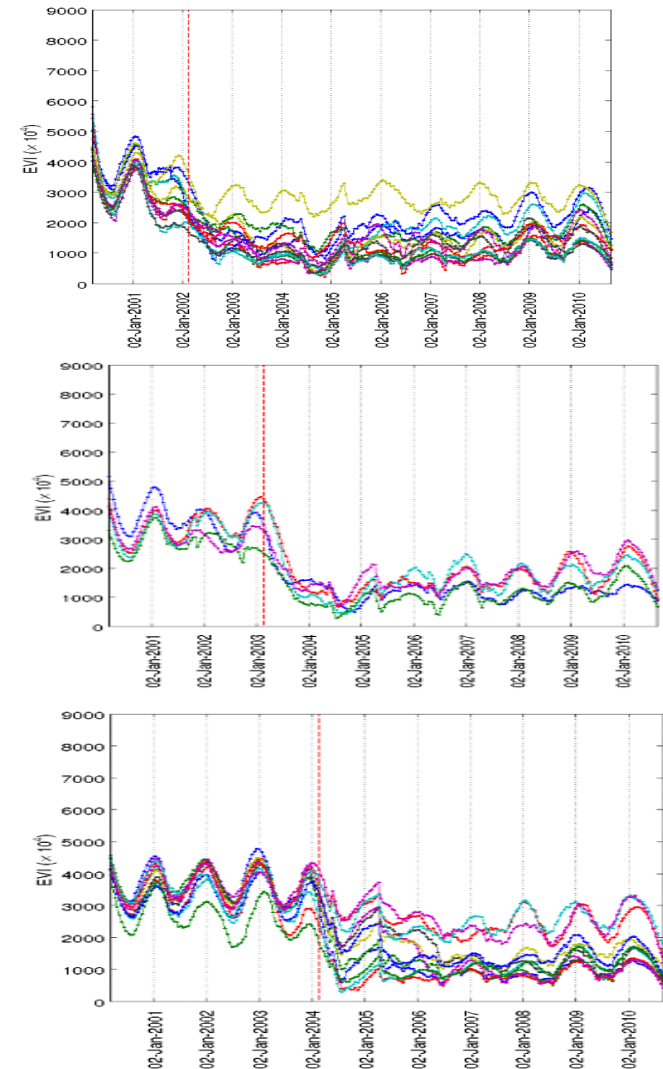
Deforestation in the Amazon Rainforest: Comparison with PRODES



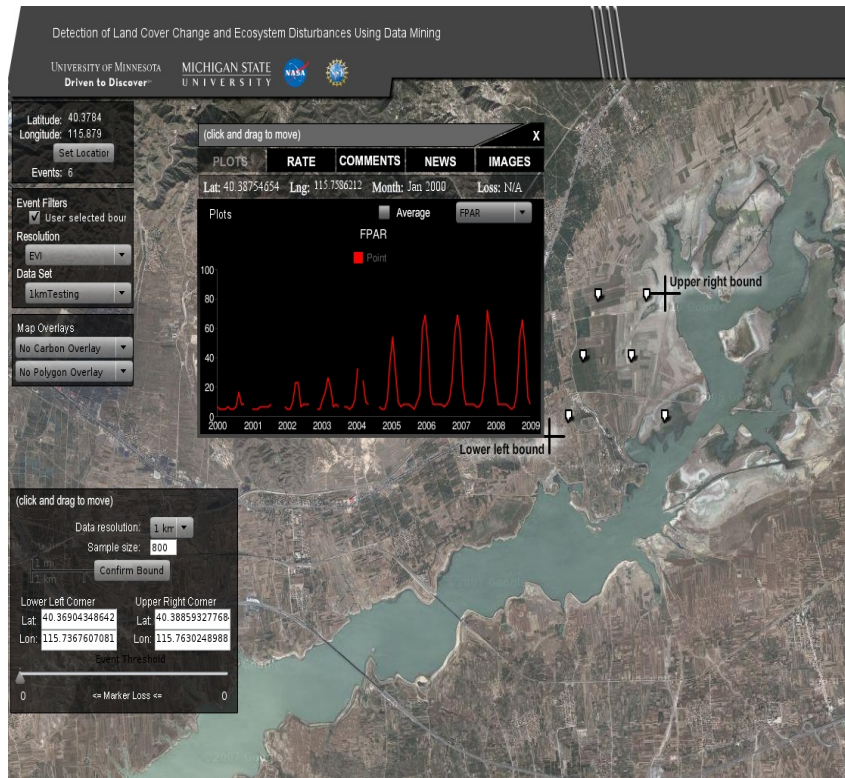
PRODES is a system for monitoring deforestation in Brazilian Amazon.

The blue polygons are deforestation changes marked by PRODES.
Yellow dots are events detected by our algorithm.

Gold Mine in Protected Forest , Tanzania



Reforestation near Guangting Reservoir, China



- These reforestation events are around Guangting Reservoir, a reservoir around 100 miles away from Beijing.

- Around 20 years ago, Guangting Reservoir used to play an important role of serving water for people in Beijing and Zhangjiakou.

- The environment around the reservoir got polluted after years, due to lack of protection.

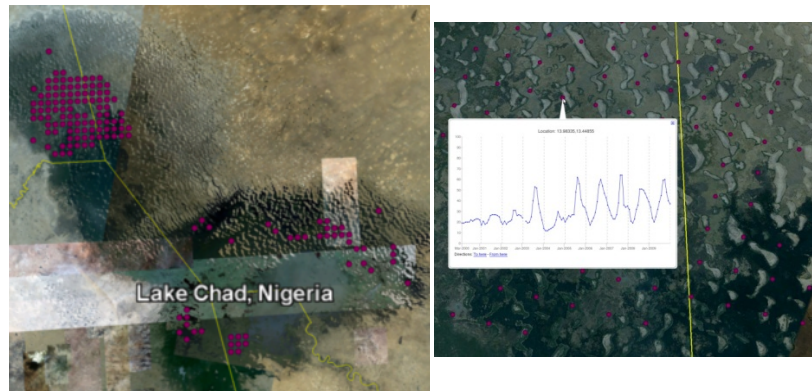
- It is located very close to Beijing and plays an important role, therefore the government began to give a comprehensive treatment for this area.

- Part of the treatment is planting trees around Guangting Reservoir which started in 2003 and is still going on.

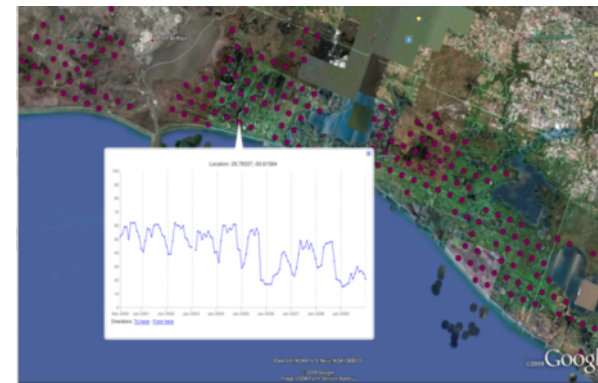
News Articles:

- http://www.yzhbw.net/news/shownews-22_510.aspx
- http://news.china.com.cn/rollnews/2010-06/04/content_2514320.htm

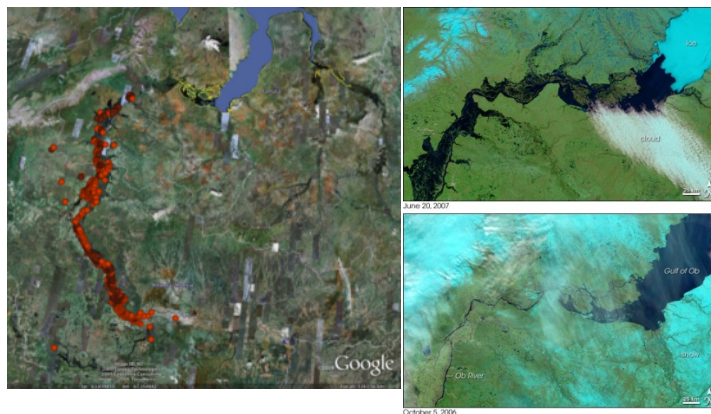
Detecting other land cover changes



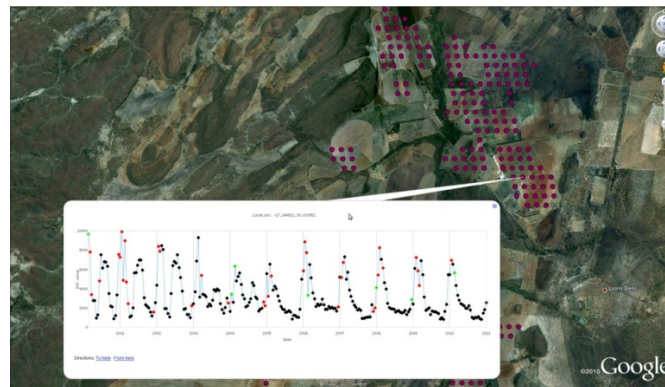
Shrinking of Lake Chad, Nigeria



Damage to vegetation by hurricane Katrina

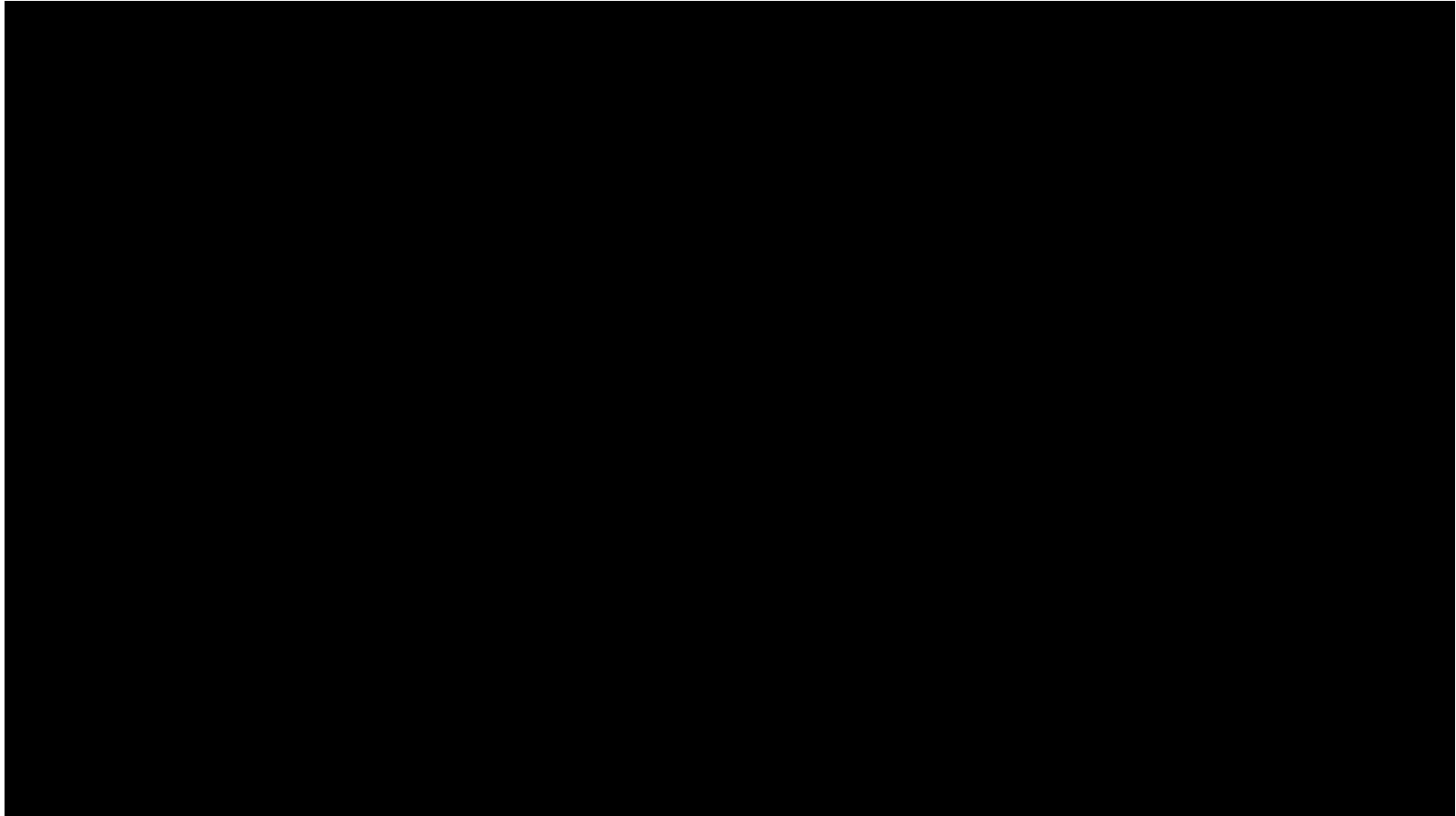


Flooding along Ob River, Russia



Farm abandonment in Zimbabwe during political conflict between 2004 and 2008.

ALERT Platform



Impact on REDD+



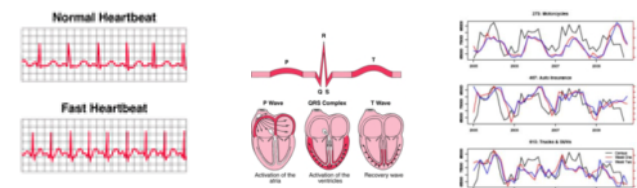
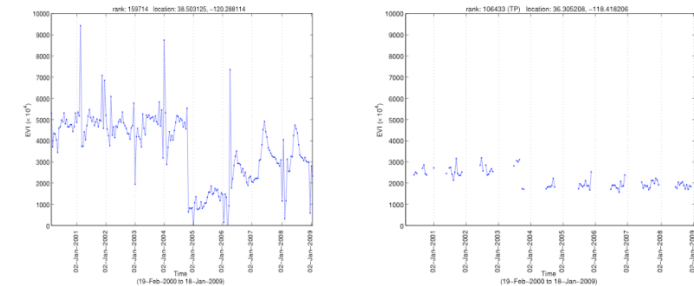
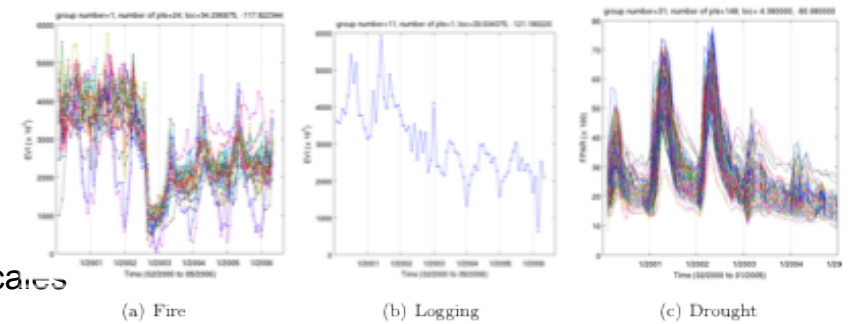
“The [Peru] government needs to spend more than \$100m a year on high-resolution satellite pictures of its billions of trees. But ... a computing facility developed by the Planetary Skin Institute (PSI) ... might help cut that budget.”

“ALERTS, which was launched at Cancún, uses ... data-mining algorithms developed at the University of Minnesota and a lot of computing power ... to spot places where land use has changed.”

- The Economist 12/16/2010

Monitoring Forest Cover Change: Challenges Ahead

- Designing robust change detection algorithms
- Characterization of land cover changes
- Multi-resolution analysis (250m vs 1km vs 4km)
 - Different kinds of changes are visible at different scales
- Multivariate analysis
 - Detecting some types of changes (e.g. crop rotations) will require additional variables.
- Data quality improvement
 - Preprocessing of data using spatio-temporal noise removal and smoothing techniques can increase performance of change detection.
- Incremental update and Real-time detection
- Spatial event identification
- Applications in variety of domains:
 - Climate, agriculture, energy
 - Economics, health care, network traffic

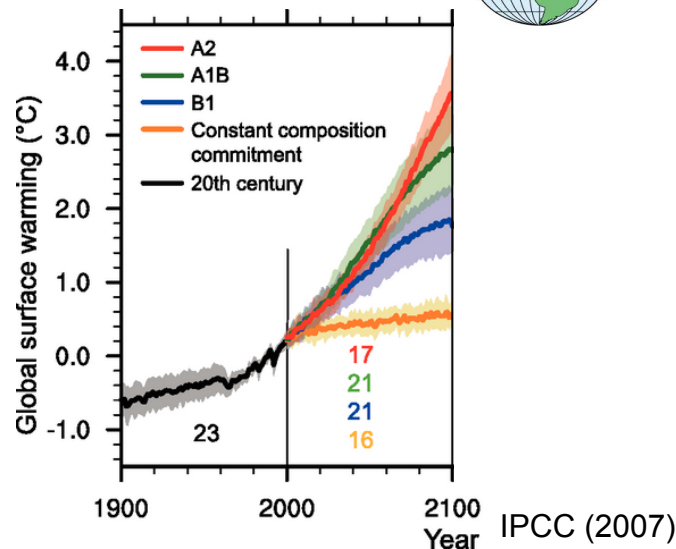
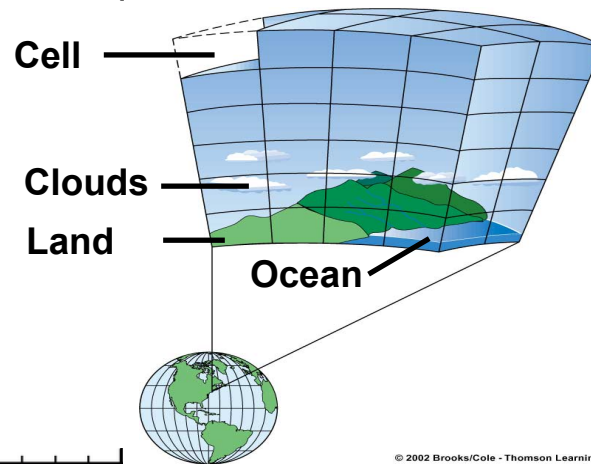


Source: Merck, Google.

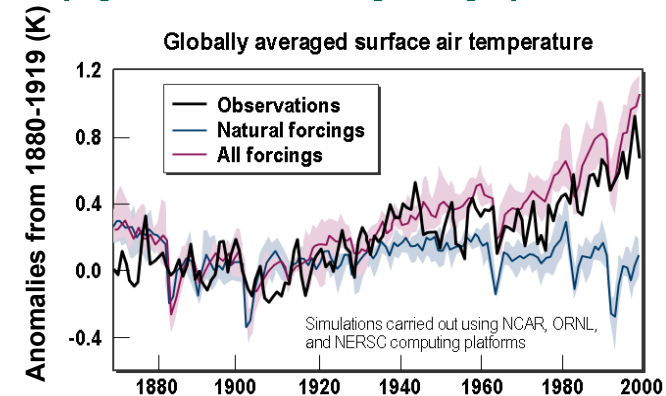
Understanding Climate Change - Physics based Approach

General Circulation Models: Mathematical models with physical equations based on fluid dynamics

Parameterization and non-linearity of differential equations are sources for uncertainty!



Temperature increases are human-induced
The anthropogenic climate change “fingerprint”



In the absence of human-induced changes to the atmosphere, the earth would be in a cooling trend

Figure Courtesy: ORNL

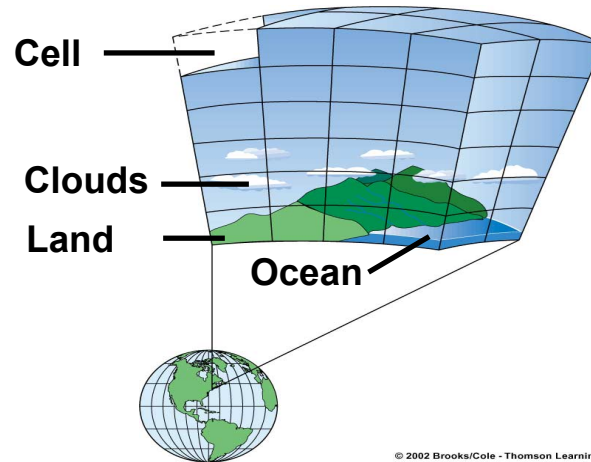
Projection of temperature increase under different **Special Report on Emissions Scenarios (SRES)** by 24 different GCM configurations from 16 research centers used in the **Intergovernmental Panel on Climate Change (IPCC) 4th Assessment Report**.

A1B: “integrated world” balance of fuels
A2: “divided world” local fuels
B1: “integrated world” environmentally conscious

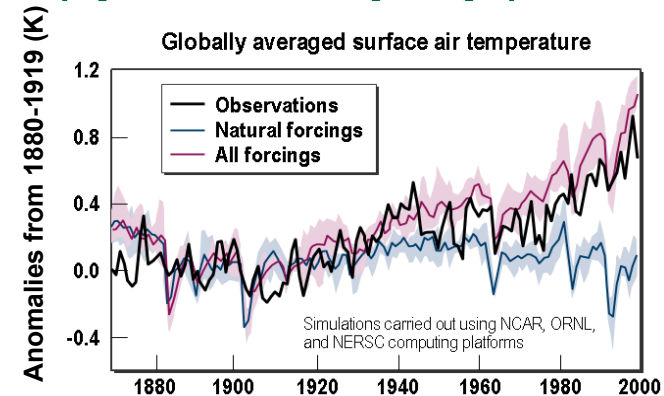
Understanding Climate Change - Physics based Approach

General Circulation Models: Mathematical models with physical equations based on fluid dynamics

Parameterization and non-linearity of differential equations are sources for uncertainty!



Temperature increases are human-induced
The anthropogenic climate change “fingerprint”



In the absence of human-induced changes to the atmosphere, the earth would be in a cooling trend

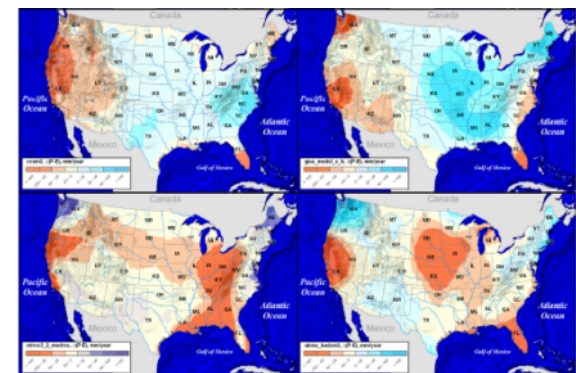
Figure Courtesy: ORNL

Physics-based models are essential but not adequate

- Relatively reliable predictions at global scale for ancillary variables such as temperature
- Least reliable predictions for variables that are crucial for impact assessment such as regional precipitation

"The sad truth of climate science is that the most crucial information is the least reliable" (Nature, 2010)

Disagreement between IPCC models



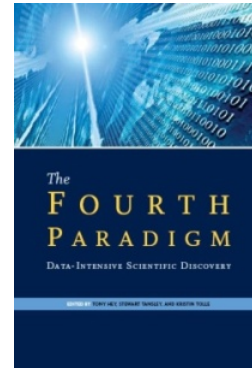
Regional hydrology exhibits large variations among major IPCC model projections

NSF Expedition: Understanding Climate Change - A *Data-Driven Approach*



Project aim:

A new and transformative data-driven approach that complements physics-based models and improves prediction of the potential impacts of climate change



"... data-intensive science [is] ...a new, fourth paradigm for scientific exploration." - Jim Gray

Transformative Computer Science Research

Predictive Modeling

Enable predictive modeling of typical and extreme behavior from multivariate spatio-temporal data

Complex Networks

Enable studying of collective behavior of interacting eco-climate systems

Relationship Mining

Enable discovery of complex dependence structures: non-linear associations or long range spatial dependencies

High Performance Computing

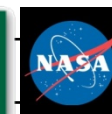
Enable efficient large-scale spatio-temporal analytics on exascale HPC platforms with complex memory hierarchies

• Science Contributions

- Data-guided uncertainty reduction by blending physics models and data analytics
- A new understanding of the complex nature of the Earth system and mechanisms contributing to adverse consequences of climate change

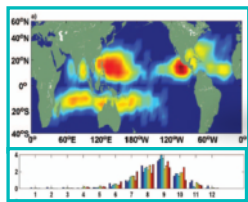
• Success Metric

- Inclusion of data-driven analysis as a standard part of climate projections and impact assessment (e.g., for IPCC)

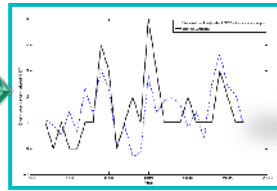


Some Driving Use Cases

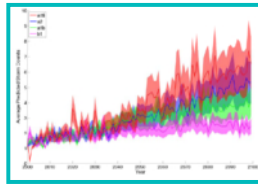
Impact of Global Warming on Hurricane Frequency



Find non-linear relationships



Validate w/ hindcasts

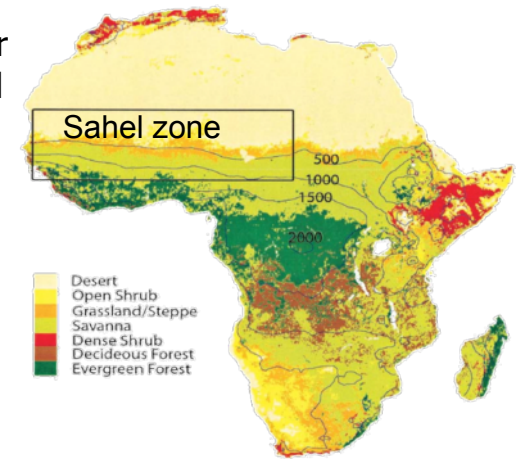


Build hurricane models

Regime Shift in Sahel

Onset of major 30-year drought over the Sahel region in 1969

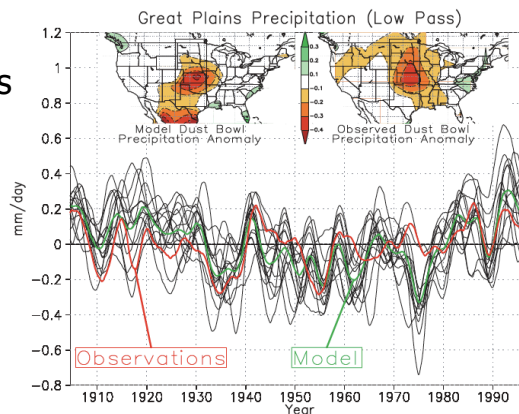
Regime shift can occur without any advanced warning and may be triggered by isolated events such as storms, drought



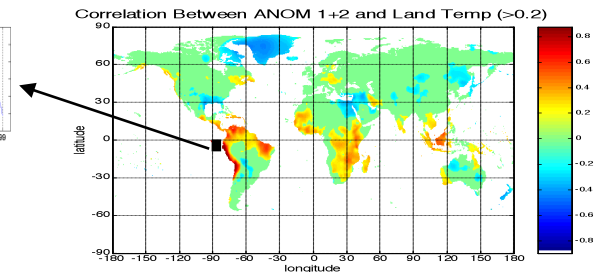
1930s Dust Bowl

Affected almost two-thirds of the U.S. Centered over the agriculturally productive Great Plains

Drought initiated by anomalous tropical SSTs (Teleconnections)

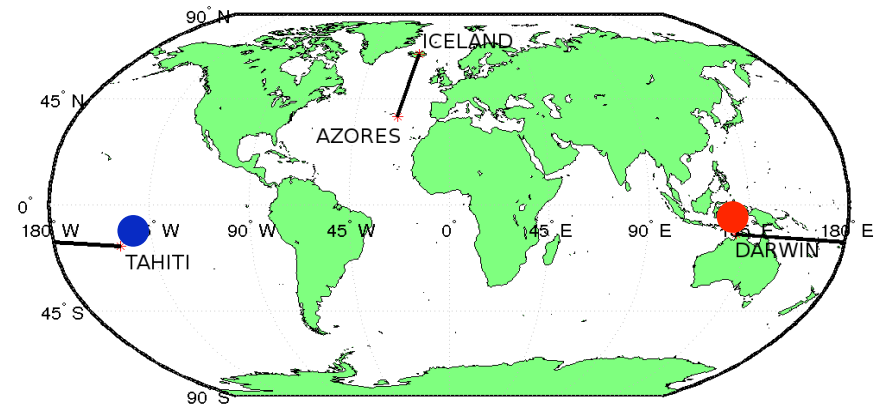
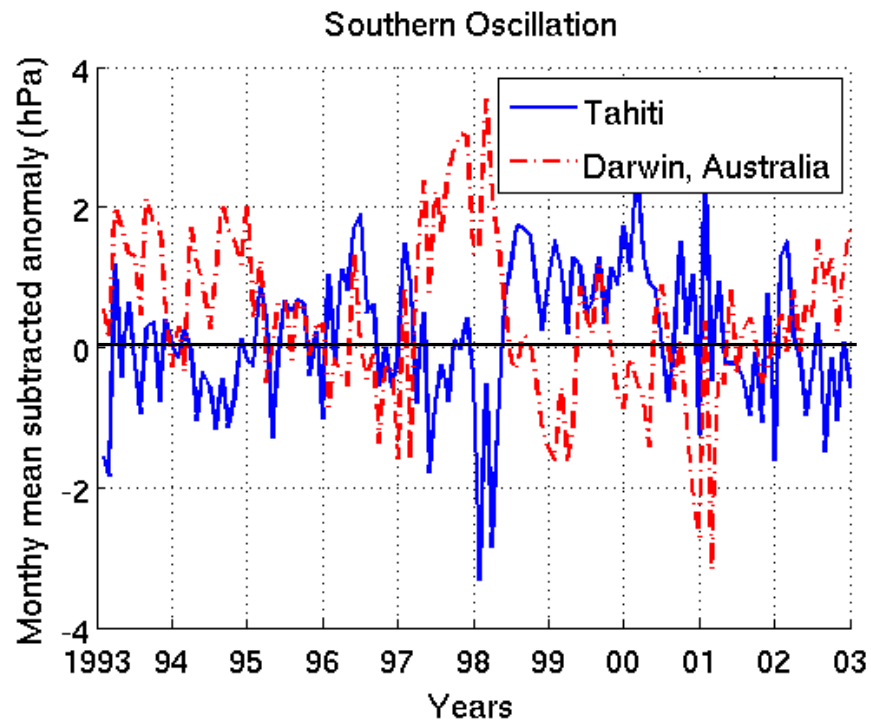


Discovering Climate Teleconnections



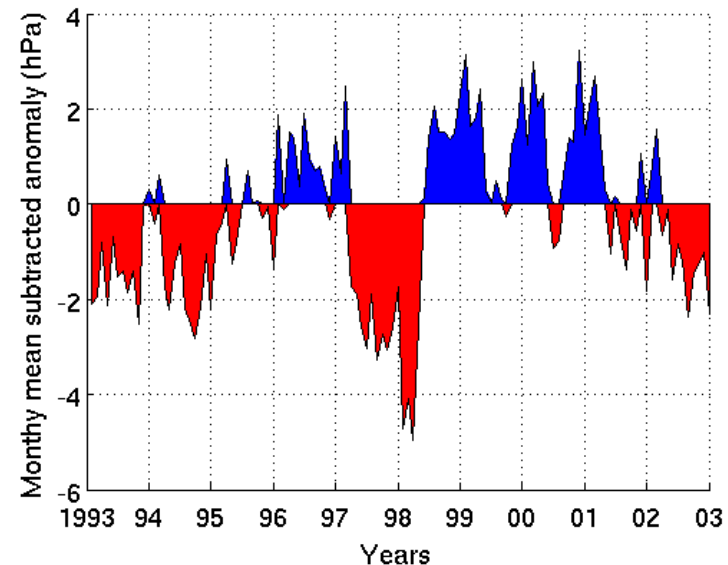
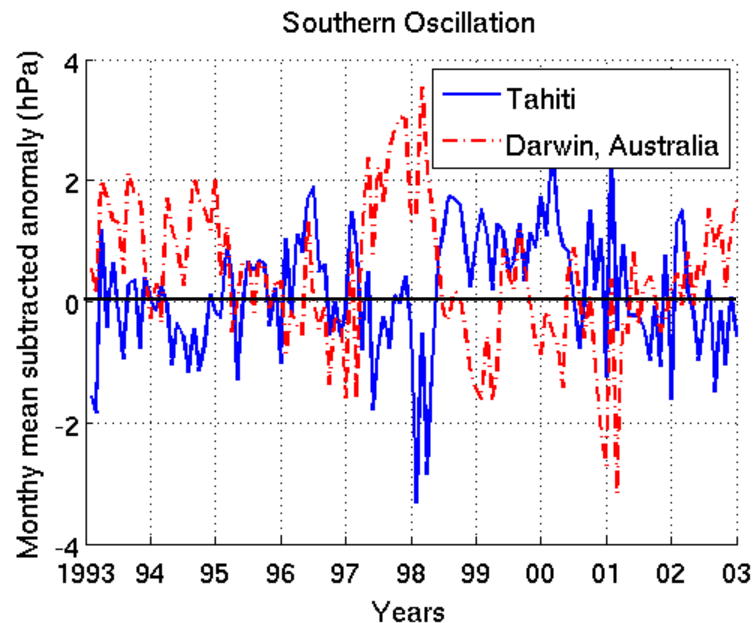
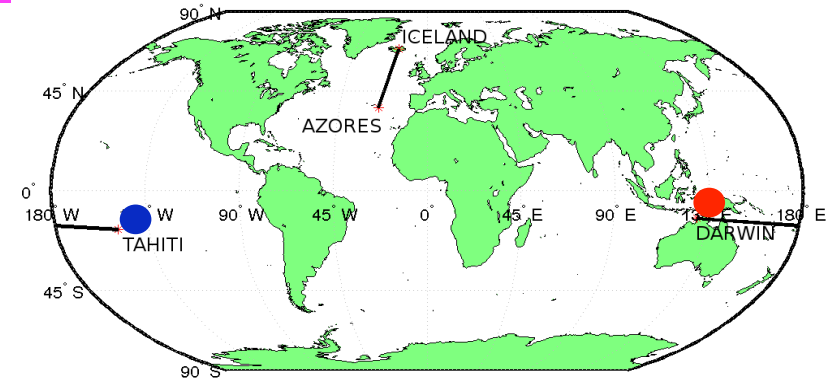
Understanding climate variability using Dipole Analysis

Dipoles represent a class of teleconnections characterized by anomalies of opposite polarity at two locations at the same time.



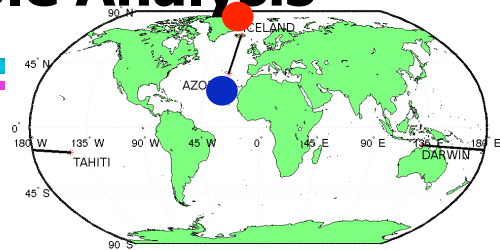
Understanding climate variability using Dipole Analysis

Dipoles represent a class of teleconnections characterized by anomalies of opposite polarity at two locations at the same time.

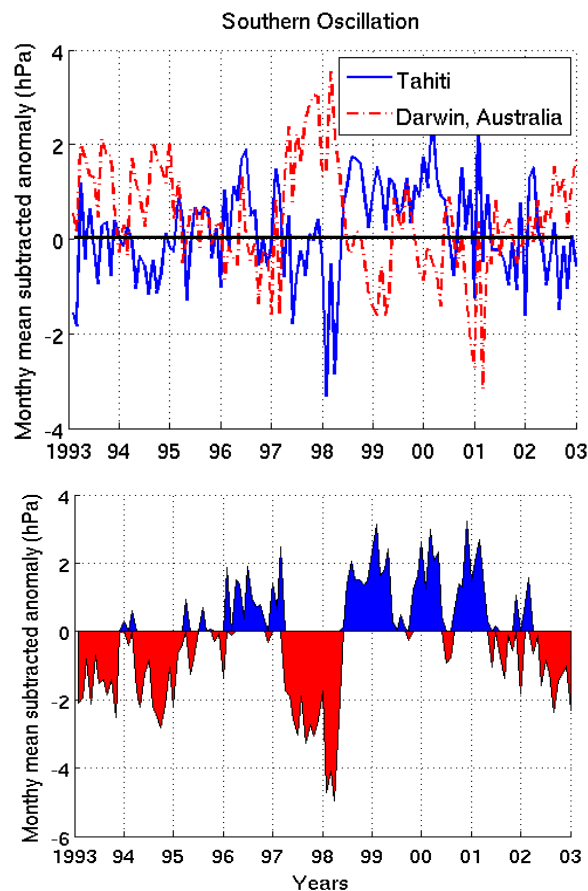


Understanding climate variability using Dipole Analysis

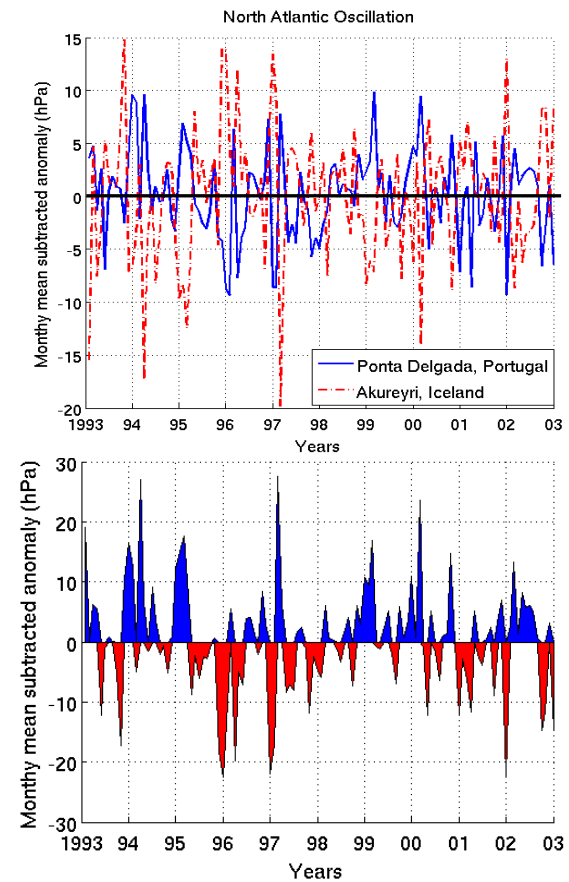
Dipoles represent a class of teleconnections characterized by anomalies of opposite polarity at two locations at the same time.



Southern Oscillation: Tahiti and Darwin



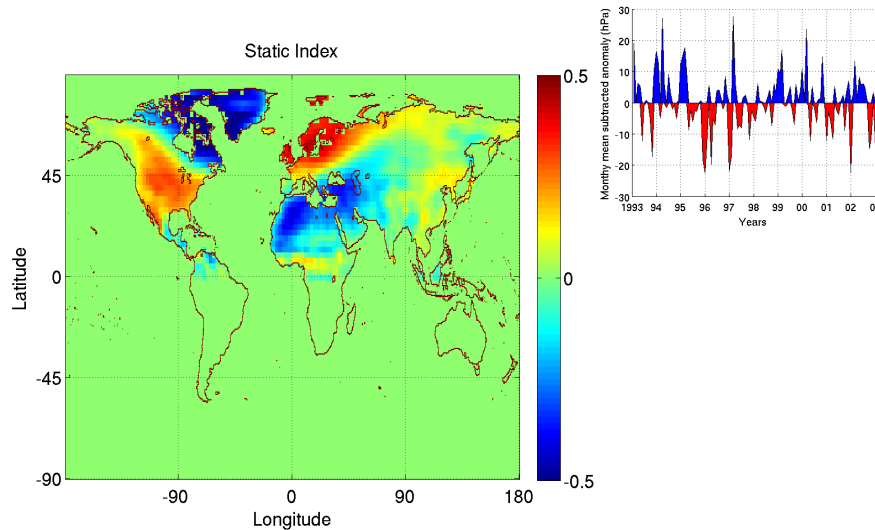
North Atlantic Oscillation: Iceland and Azores



Importance of dipoles

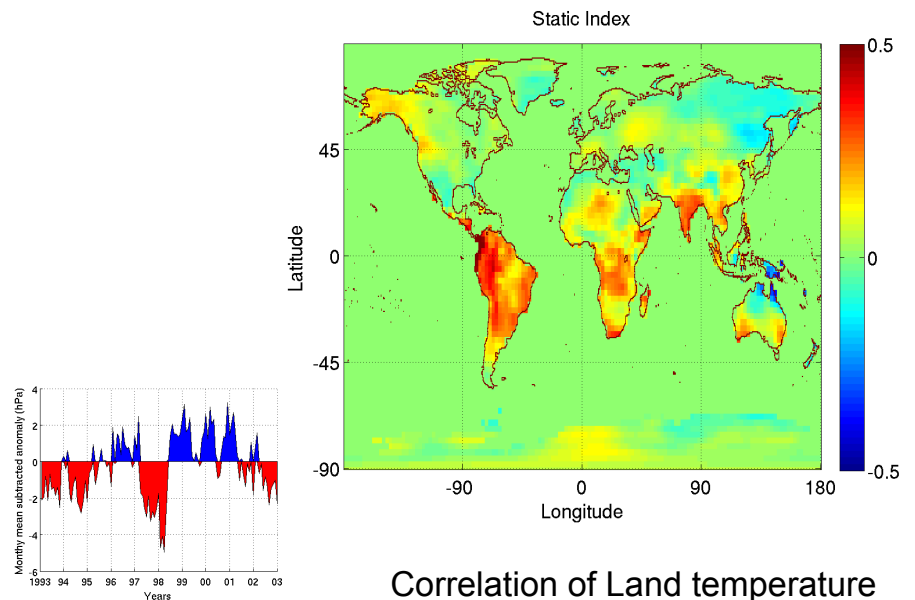
Crucial for understanding the climate system, especially for weather and climate forecast simulations within the context of global climate change.

NAO influences sea level pressure (SLP) over most of the Northern Hemisphere. Strong positive NAO phase (strong Icelandic Low and strong Azores High) are associated with above-average temperatures in the eastern US.










Correlation of Land temperature anomalies with NAO

SOI dominates tropical climate with floodings over East Asia and Australia, and droughts over America. Also has influence on global climate.



Correlation of Land temperature anomalies with SOI

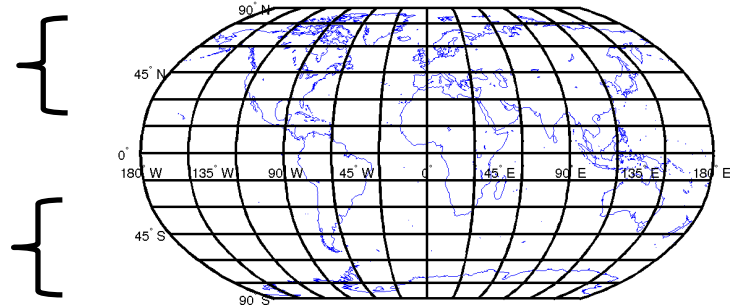
List of Well Known Climate Indices

Index	Description
SOI	 Southern Oscillation Index: Measures the SLP anomalies between Darwin and Tahiti
NAO	 North Atlantic Oscillation: Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	 Arctic Oscillation: Defined as the first principal component of SLP northward of 20° N
PDO	 Pacific Decadal Oscillation: Derived as the leading principal component of monthly SST anomalies in the North Pacific Ocean, poleward of 20° N
WP	 Western Pacific: Represents a low-frequency temporal function of the 'zonal dipole' SLP spatial pattern involving the Kamchatka Peninsula, southeastern Asia and far western tropical and subtropical North Pacific
PNA	 Pacific North American: SLP Anomalies over the North Pacific Ocean and the North America
AAO	 Antarctic Oscillation: Defined as the first principal component of SLP southward of 20° S
NINO1+2	Sea surface temperature anomalies in the region bounded by 80° W-90° W and 0° -10° S
NINO3	Sea surface temperature anomalies in the region bounded by 90° W-150° W and 5° S-5° N
NINO3.4	Sea surface temperature anomalies in the region bounded by 120° W-170° W and 5° S-5° N
NINO4	Sea surface temperature anomalies in the region bounded by 150° W-160° W and 5° S-5° N

Discovered primarily by human observation and EOF Analysis

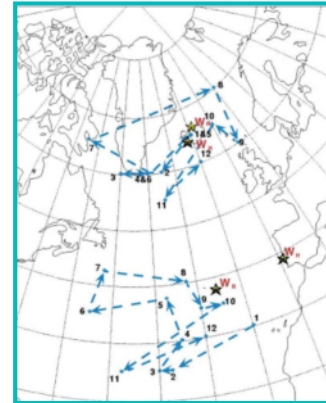
AO: EOF Analysis of 20N-90N Latitude

AAO: EOF Analysis of 20S-90S Latitude



Motivation for Automatic Discovery of Dipoles

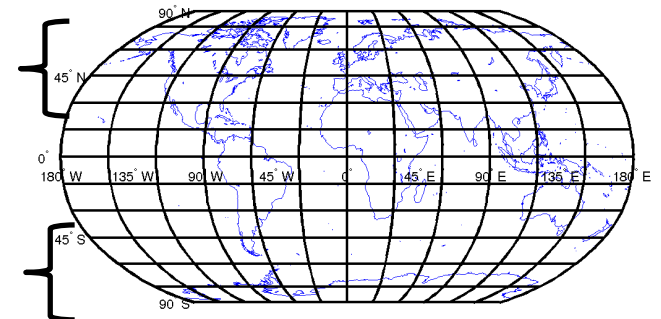
- The known dipoles are defined by static locations but the underlying phenomenon is dynamic
- Manual discovery can miss many dipoles
- EOF and other types of eigenvector analysis finds the strongest signals and the physical interpretation of those can be difficult.



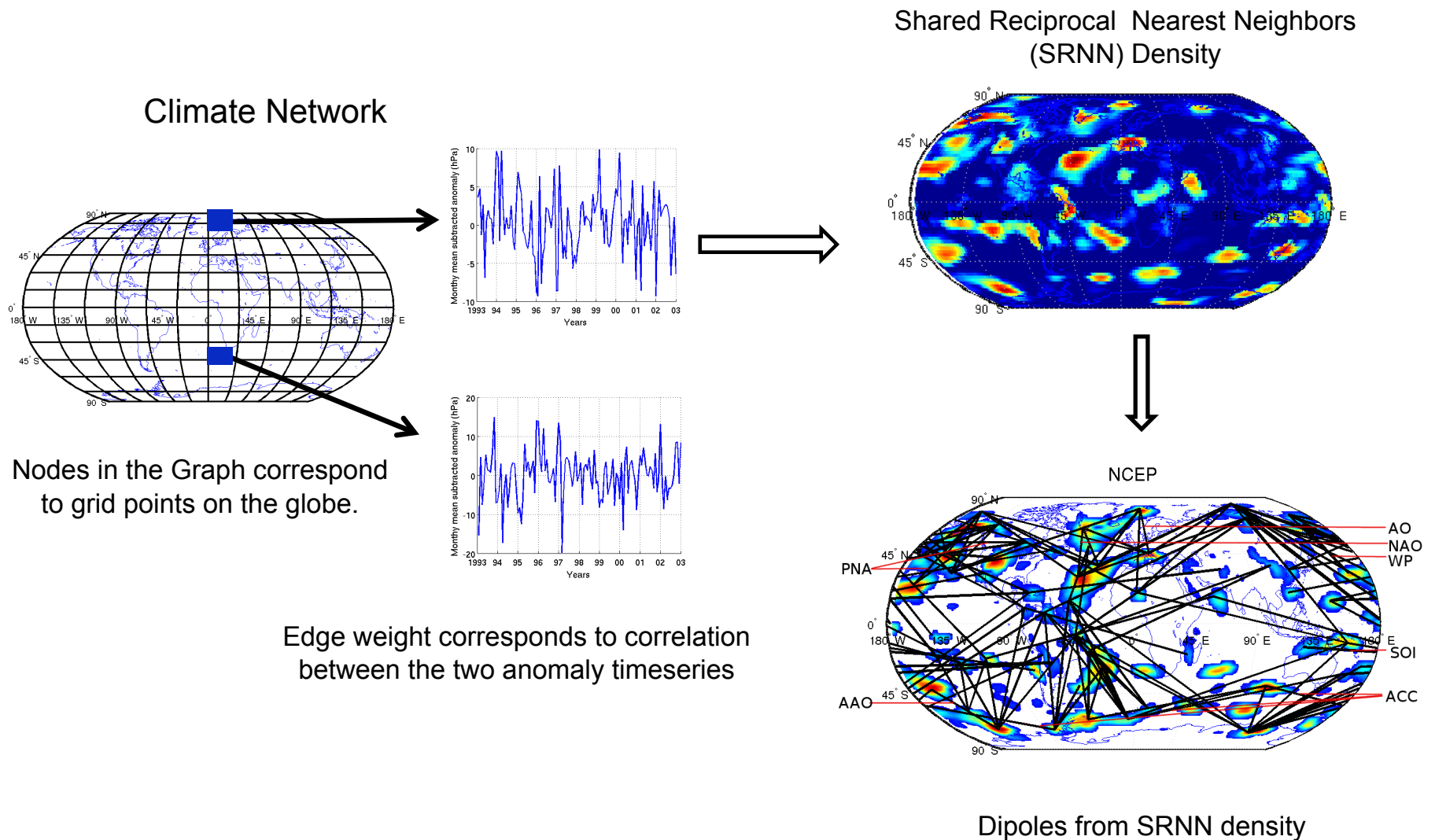
Dynamic behavior of the high and low pressure fields corresponding to NOA climate index (Portis et al, 2001)

AO: EOF
Analysis of
20N-90N Latitude

AAO: EOF
Analysis of
20S-90S Latitude



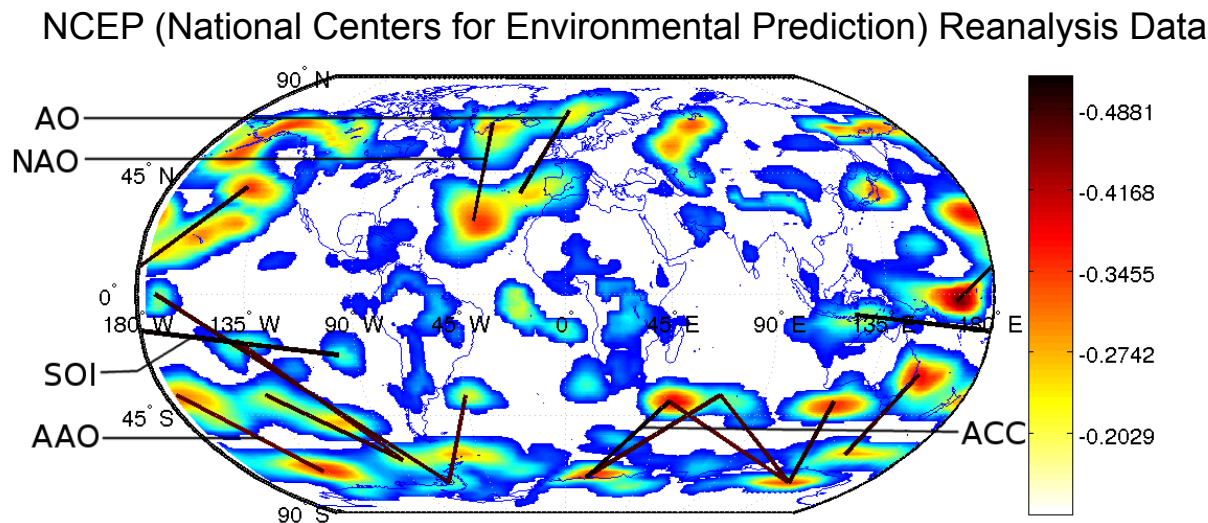
Discovering Climate Teleconnections using Network Representation



Benefits of Automatic Dipole Discovery

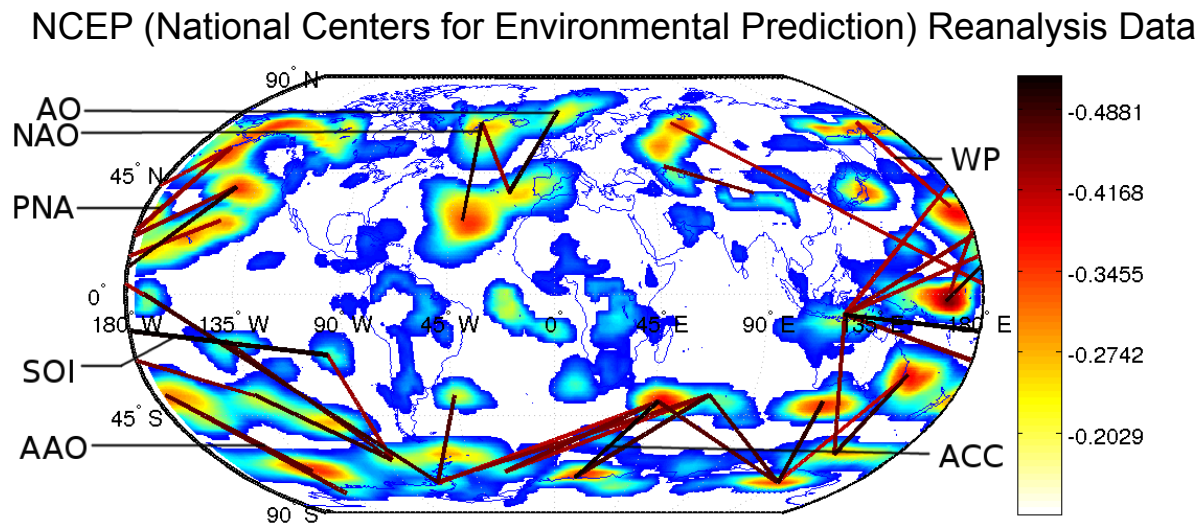
- Detection of Global Dipole Structure
 - Most known dipoles discovered
 - New dipoles may represent previously unknown phenomenon.
 - Enables analysis of relationships between different dipoles
- Location based definition possible for some known indices that are defined using EOF analysis.
- Dynamic versions are often better than static
- Dipole structure provides an alternate method to analyze GCM performance

Detection of Global Dipole Structure



- Most known dipoles discovered
- Location based definition possible for some known indices that are defined using EOF analysis.
- New dipoles may represent previously unknown phenomenon.

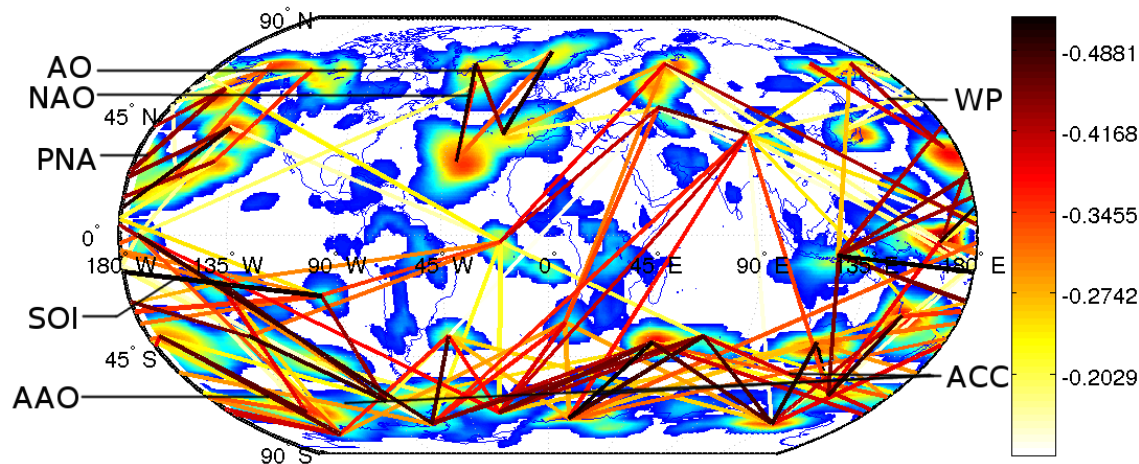
Detection of Global Dipole Structure



- Most known dipoles discovered
- Location based definition possible for some known indices that are defined using EOF analysis.
- New dipoles may represent previously unknown phenomenon.

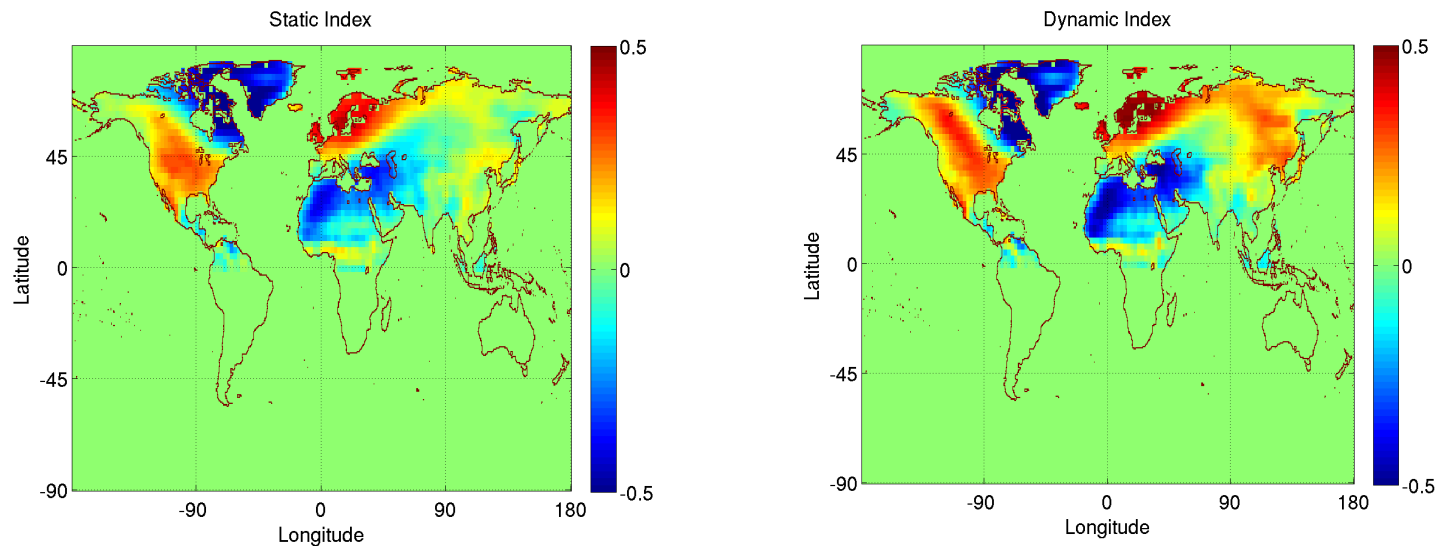
Detection of Global Dipole Structure

NCEP (National Centers for Environmental Prediction) Reanalysis Data



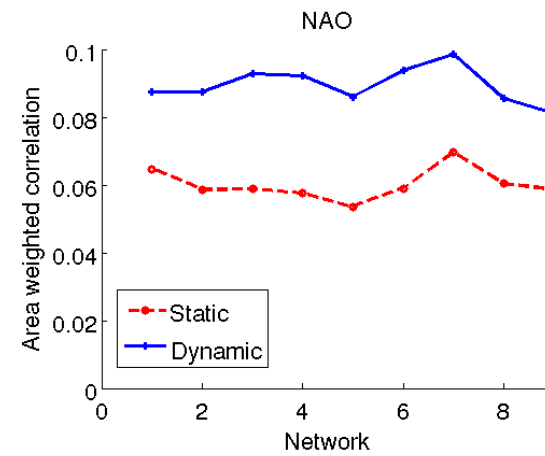
- Most known dipoles discovered
- Location based definition possible for some known indices that are defined using EOF analysis.
- New dipoles may represent previously unknown phenomenon.

Static vs Dynamic NAO Index: Impact on land temperature

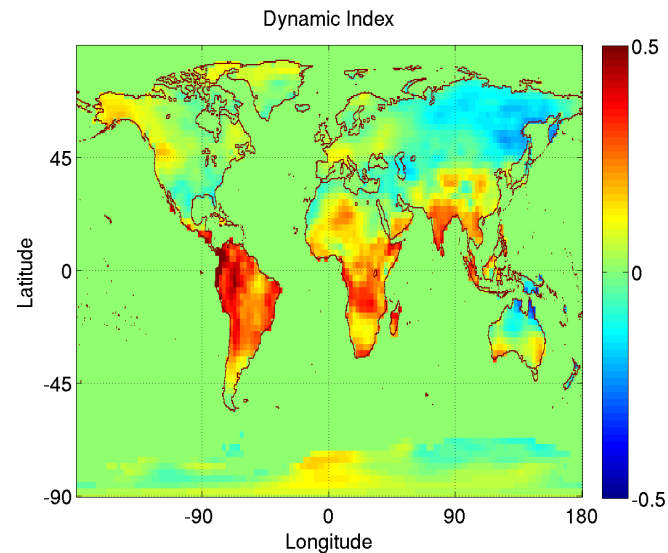
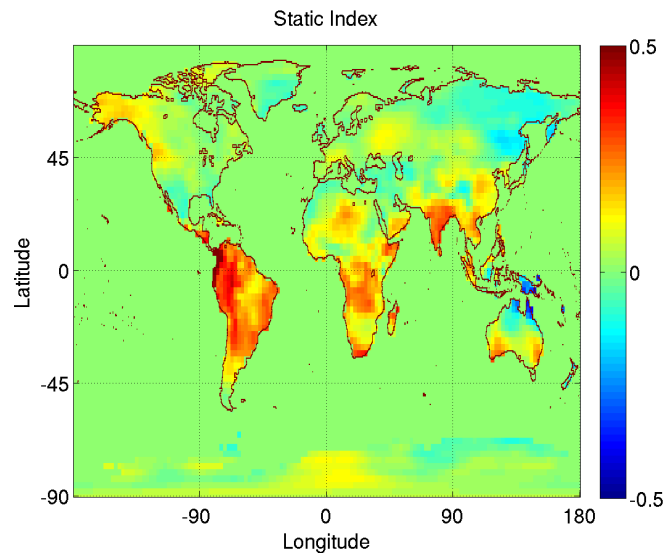


The dynamic index generates a stronger impact on land temperature anomalies as compared to the static index.

Figure to the right shows the aggregate area weighted correlation for networks computed for different 20 year periods during 1948-2008.

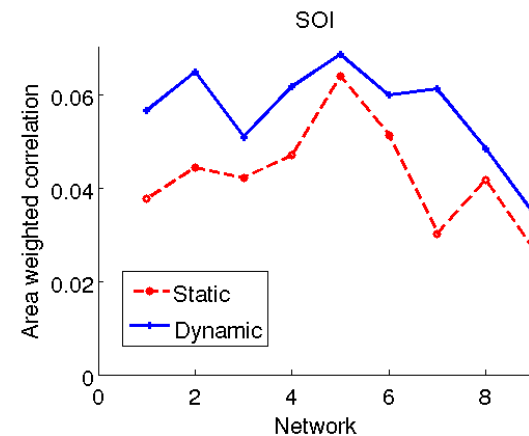


Static vs Dynamic SO Index: Impact on land temperature



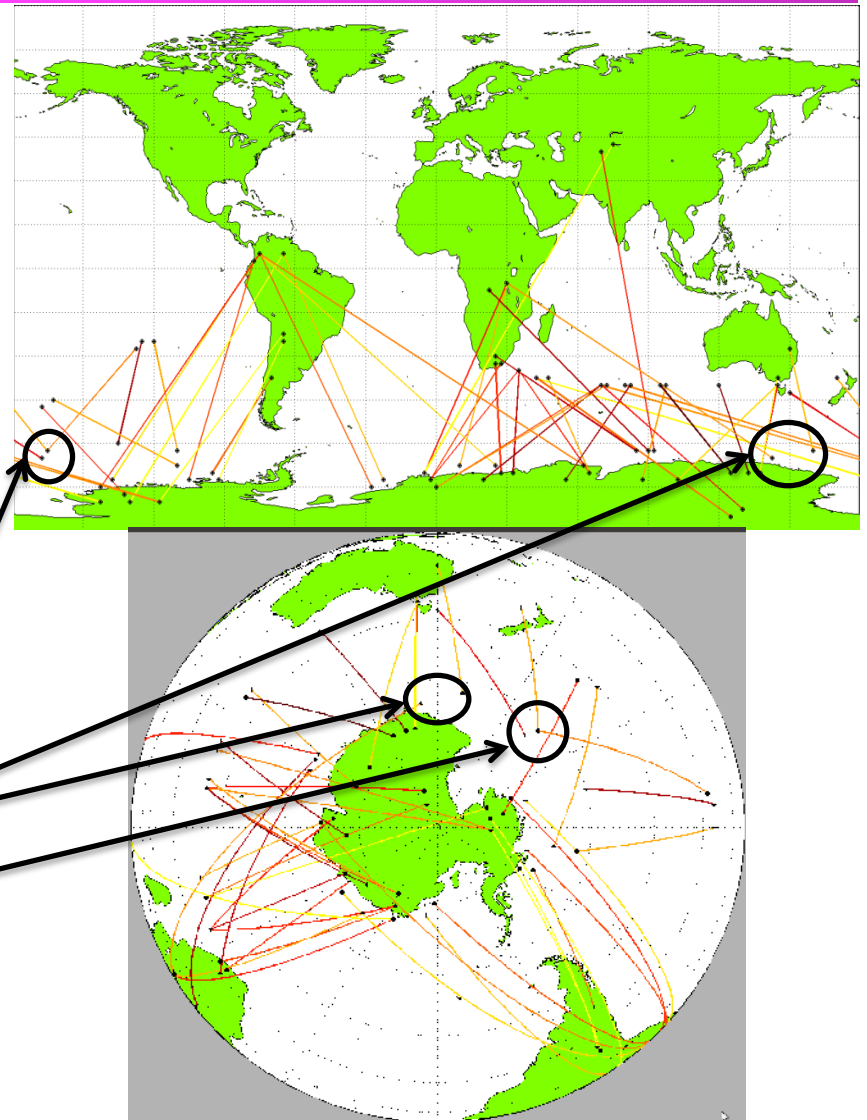
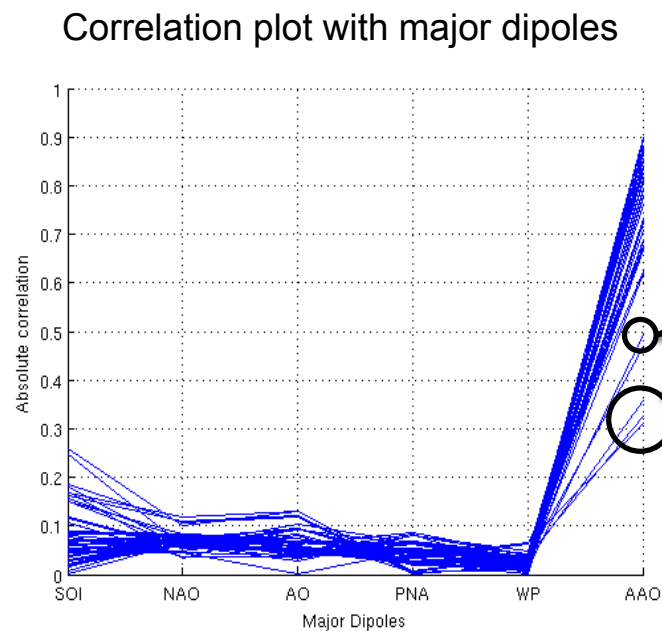
The dynamic index generates a stronger impact on land temperature anomalies as compared to the static index.

Figure to the right shows the aggregate area weighted correlation for networks computed for different 20 year periods during 1948-2008.



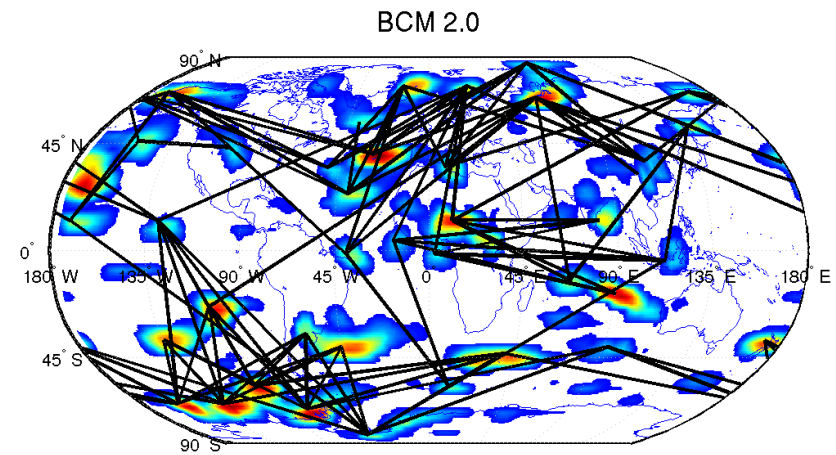
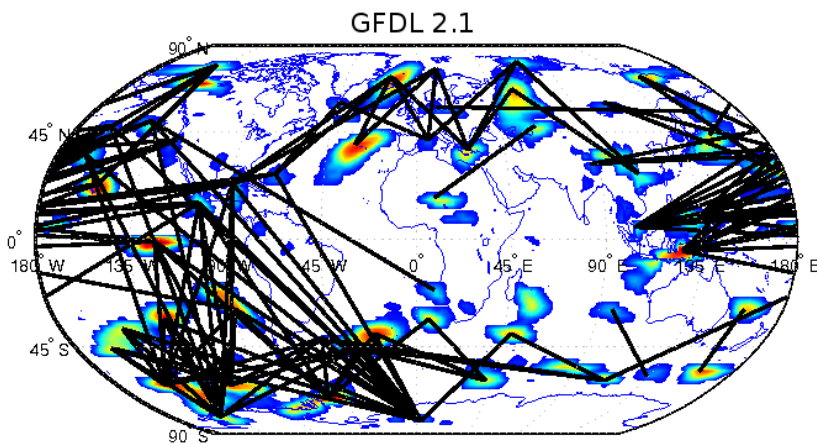
A New Dipole Around Antarctica?

- 3 major dipole structures can be seen.
- The AAO and two others shown in figure
- A newer phenomenon which is not captured by the EOF analysis?



Comparison of Climate Models using Dipole Structure

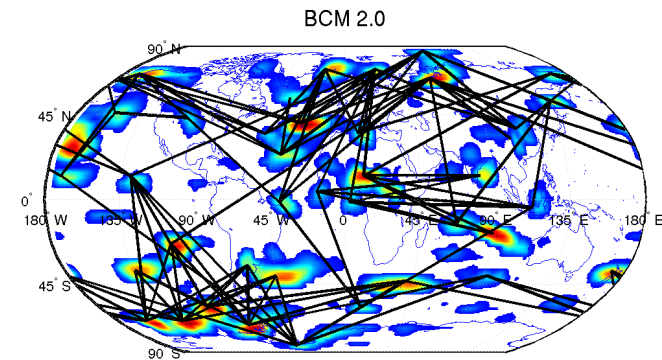
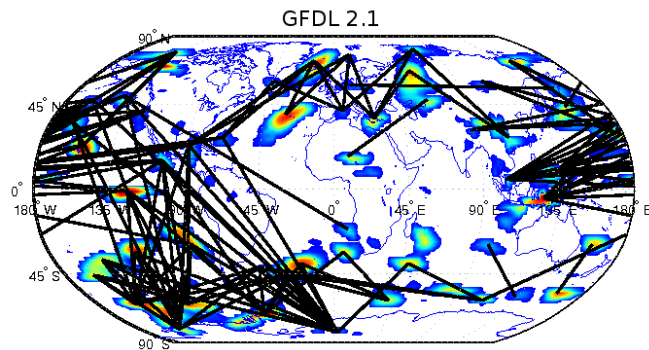
Hindcast data



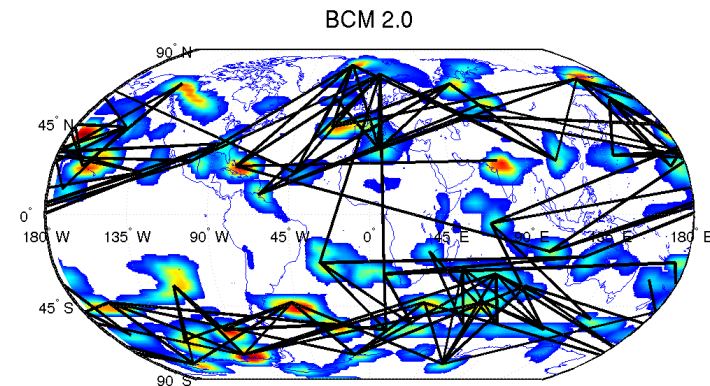
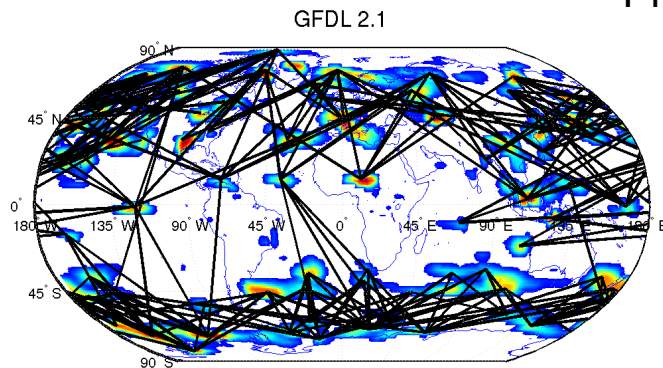
- Differences in dipole structure can offer valuable insights to climate scientists on model performance
- Strength of the dipoles varies in different climate models
 - SOI is only simulated by GFDL 2.1 and not by BCM 2.0.

Comparison of Climate Models using Dipole Structure

Hindcast data



Projection data

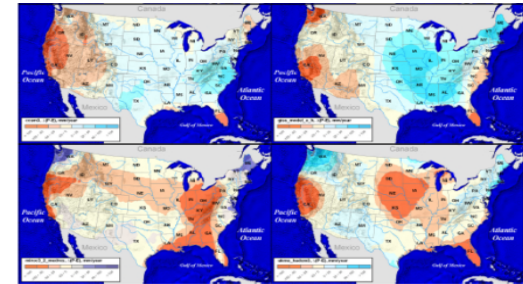


- Dipole connections in forecast data provide insights about dipole activity in future.
- For e.g. both forecasts for 2080-2100 show continuing dipole activity in the extratropics but decreased activity in the tropics. SOI activity is reduced in GFDL2.1 and activity over Africa is reduced in BCM 2.0. This is consistent with archaeological data from 3 mil. years ago, when climate was 2-3°C warmer (Shukla, et. al).

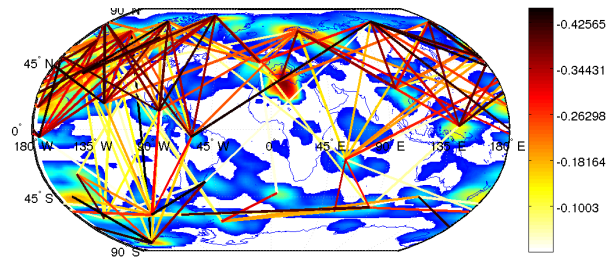
Relating Dipole Structure to Model Prediction

- The dipole structure of the top 2 models are different from the bottom two models
 - NCAR-CCSM and NASA-GISS miss SOI and other dipoles near the Equator

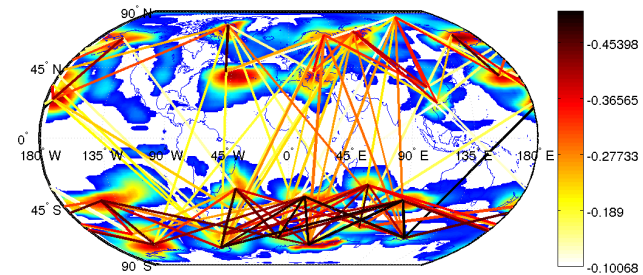
Disagreement between IPCC models



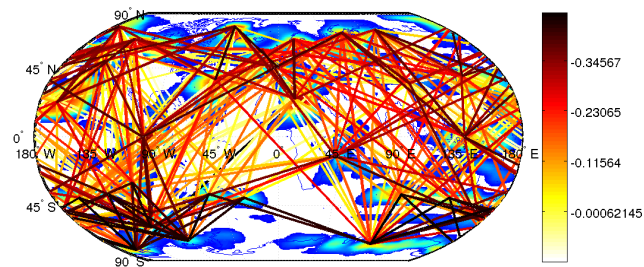
NCAR-CCSM



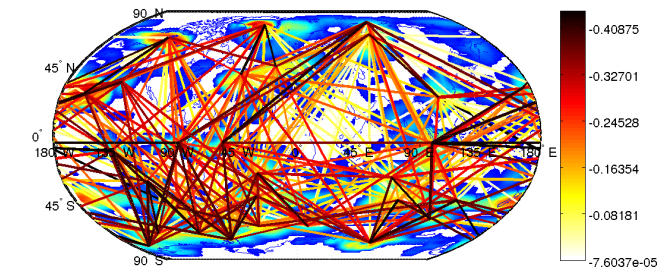
NASA-GISS



MIROC_3_2

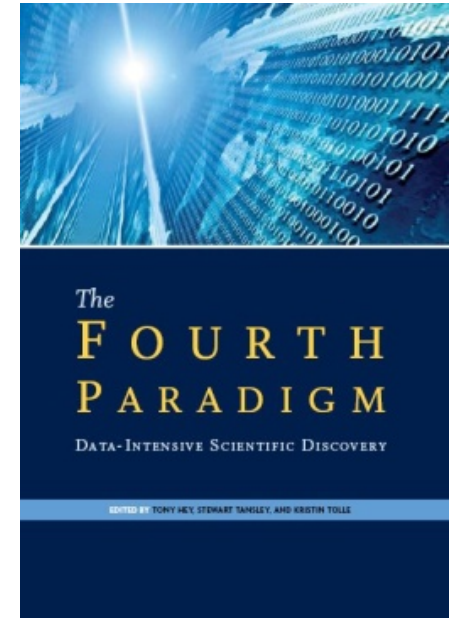


UKMO-HadCM3



Summary

- Data driven discovery methods hold great promise for advancement in the mining of climate and ecosystem data.
- Scale and nature of the data offer numerous challenges and opportunities for research in mining large datasets.



"The world of science has changed ... data-intensive science [is] so different that it is worth distinguishing [it] ... as a new, fourth paradigm for scientific exploration." - Jim Gray

Team Members and Collaborators

Michael Steinbach, Shyam
Boriah, Rohit Gupta, Gang
Fang, Gowtham Atluri, Varun
Mithal, Ashish Garg, Vanja
Paunic, Sanjoy Dey, Jaya
Kawale, Marc Dunham, Divya
Alla, Ivan Brugere, Vikrant
Krishna, Yashu Chamber, Xi
Chen, James Faghmous,
Arjun Kumar, Stefan Liess

Sudipto Banerjee, Chris Potter,
Fred Semazzi, Nagiza Samatova,
Steve Klooster, Auroop Ganguly,
Pang-Ning Tan, Joe Knight,
Arindam Banerjee, Peter Snyder

Project website

Climate and Eco-system: www.cs.umn.edu/~kumar/nasa-umn